THE UNIVERSITY *of* EDINBURGH

A feasibility study exploring the spatial distribution of *Plasmodium falciparum*

Carol Wangui Hunja

Thesis submitted for a degree of Doctor of Philosophy (PhD)

University of Edinburgh

October, 2012

Declaration

I hereby declare that except the contributions specified in the following sections, the work presented here was planned, performed and written by myself only and has not been submitted for any other degree or professional qualification.

Carol Wangui Hunja

October, 2012

Table of Contents

List of Figures

List of Tables

Dedication

This thesis is dedicated to Prof. Kenneth Murray, founder of

The Darwin Trust of Edinburgh

List of contributors

1. The *Plasmodium falciparum* artificial clone mixtures at different proportions used in Chapter 4 analysis were prepared by Prof. Richard Carter.

2. The following markers were developed by Dr. Sandra Cheesman – Pf-01-393153, Pf-01-411796, Pf-03-138-7, Pf-03-350224, Pf-03-613146, Pf-03-842383, Pf-03-967838, Pf-10-131053 and Pf-14-2085411.

3. Mali isolates were provided to us by collaboration with Dr. Ogobara Doumbo's group in Mali through links established by Dr. Teun Bousema. Cameroon isolates were provided to us by Dr. Damian Anong from University of Buea, Cameroon. The Kenya isolates were collected with tremendous help from Dr. Richard Mukabana. These isolates were used in the analysis described in Chapters 5 & 6.

4. The microsatellite analysis described in Chapters 5 & 6 was done by Dr. Richard Culleton from Nagasaki University, Japan.

Acknowledgements

Firstly I thank God for His Love and continual blessings. He is a wellspring of guidance and strength to me.

My sincerest gratitude goes to The Darwin Trust of Edinburgh and the Wellcome Trust for funding my PhD.

I would like to express my sincerest appreciation to Prof. Richard Carter my supervisor who was patient, kind and was most of all, a mentor. He not only taught me how to do good research, how to write a good thesis but also helped shape my character.

Thanks to Dr. Sandra Cheesman my second supervisor who developed some of the assays I used and also gave me support and advise whenever things were tough for me.

To Prof. Graham Stone and Dr. Joanne Thompson for stepping in as my supervisors, thank you very much.

To Mona Al Dossari who began this project. I owe a lot to your organisation which made my taking over the project much simpler.

To Kathryn Degnan, who helped me settle in the lab and taught me all that I know. Thanks hen.

I would like to express my sincerest gratitude to those who contributed to sample collections: Dr. Richard Mukabana - Rusinga Island, Kenya; Dr. Damian Anong - Buea, Cameroon; Dr. Teun Bousema who established the connection with Dr. Ogobara Doumbo - Kolle, Mali and, to the participants who consented to be a part of this study.

Am grateful to Prof. Richard Culleton for his enthusiasm and hard work. The microsatellite data has proved very useful in my project.

For the data analysis, my sincerest gratitude goes to Dr. Sara Knott – statistical analysis, Dr. Deborah Charlesworth - SNP analysis. Dr. Richard Ennos – population genetics analysis, Dr. Roberta Bergero – genetics & Dr. Sam Thumbi - maps using ARCGIS software.

My thanks go to Prof. Alex Rowe's group especially Ahmed who helped me with establishing cultures for field isolates and the rest of the group for their willingness to help. I also enjoyed sneaking into your lab meetings whenever I could.

Thanks to Dr. Alison Creasey for teaching me how to establish laboratory cultures and seeing me through that part of the project.

To Dr. Paul Hunt who always listened attentively to my whining, hugged me when I felt I couldn't go on and gave me overwhelming support in spite of being very annoying. Thanks for being my daddy.

Thank you to everyone I shared an office with and who made my stay in Edinburgh amazingly wonderful, Naomi for taking me to the Pollock halls to collect my room keys (I remember that), Elaine for the mince pies (you know what I mean), Sittiporn for the wonderful cooking, Kasia for going through this with me, Les and your wonderful work in the 'moose hoose', Laurence, Laura, Matt, Achim and everyone else who has gone through GA53. We had a great time!!!

A big thank you to Prof. Aubrey Manning. I still remember his words to me 'write a good thesis Carol, it's the only one you'll ever write in your lifetime.' Well, I did my best.

To the lunch room crew – Laurence, Roberta, Irene, Richard Ennos, Nadine, JC and many others who joined us. We had amazing and stimulating conversations.

To the group from Portugal, all Drs. now. Sofia B - you gave me hope when all was lost, Louise - oh what wonderful laughs we shared, Pedro Cravo and Axel Martinelli. To you all I say obrigado/a!!

To the Orange Wednesday group. Chinese or Nandos? I had a fabulous time with you all. Ahsanteni sana!!

To Peg, Rob and Harry, happy wonderful times we shared. Thank you very much.

To my comrades from all over Africa. Thank you so much for keeping me grounded and reminding me that east or west, my home is in Africa. That's where my heart is.

To my parents Mr. James Hunja and Mrs. Rose Hunja. Thank you so much for your support. To my brother Tom and sister Loise, you have always loved me even when I was an annoying big sis and didn't deserve it. Am lucky to have you in my life

To everyone who made this journey possible but I forgot to mention, please know that I am deeply grateful and would have never made it without you. Natoa shukrani kwenu wote na mungu awabariki!!!

List of abbreviations

AEIRs - Annual Entomological Inoculation Rates

ACD - Acid Citrate Dextrose

ACTs - Artemisinin-based Combination Therapy

AFLPs - Amplified Fragment Length Polymorphisms

AMA1 - Apical Membrane Antigen 1

ATP – Adenosine Triphosphate

CCD - Charge-coupled Device

CDC - Centers for Disease Control and Prevention

CSP - Circumsporozoite Surface Protein

DSS - Demographic Surveillance System

DNA – Deoxyribonucleic Acid

GD – Genetic Distance

GoK – Government of Kenya

G6PD - Glucose-6-Phosphate Dehydrogenase

GR – Genetic Relatedness

GD – Genetic distance

HWE – Hardy-Weinberg equilibrium

IBD – Isolation by distance

iRBC – Infected Red Blood Cell

IRS - Indoor residual spraying
ITNs - Insecticide treated nets

LD - Linkage disequilibrium

LLNs - Long Lasting Insecticide Treated Nets

MalERA - Malaria Eradication Research Agenda

MOI – Multiplicity of infection

MSP1 & 2 - Merozoite Surface Proteins 1 & 2

MVTR - Malaria Vaccine Technology Roadmap

NCBI - National Centre for Biotechnology Information

NVBDCP - National Vector Borne Disease Control Programme

Packed Cell Volume - PCV

PBS – Phosphate Buffered Saline

PCR – Polymerase Chain Reaction

PfEMP1 – *Plasmodium falciparum* Erythrocyte Membrane Protein 1

PPi - Inorganic Pyrophosphate

pRBC – Parasitized Red Blood Cell

RAP2 - Recombinant Rhoptry-Associated Protein 2

RBC – Red Blood Cell

RFLPs - Restriction Fragment Length Polymorphisms

RPM - Revolutions per Minute

SNP – Single Nucleotide Polymorphism

TBVs - Transmission Blocking Vaccines

TMMHC - Tom Mboya Memorial Health Centre

TRAP - Thrombospondin-related Anonymous Protein

TNF-α – Tumor Necrosis Factor α

VIMT - Vaccines that Interrupt Malaria Transmission

WHO - World Health Organisation

Abstract

The way malaria parasites are transmitted in space will have an influence on their genetic relationships. It can be expected that parasites collected within close geographic distances of each other would be more closely related than those across large geographic distances. Further to this, because malaria transmission is focal and heterogeneous in space, then the genetic relatedness between malaria parasites in these foci of malaria transmission would be greater within tightly clustered regions. Thus, using the level of genetic relatedness of these parasites would reveal how they are transmitted not only within these foci but at different geographic settings. This knowledge would offer insight on how malaria control methods can be effectively disseminated.

In field settings malaria infections are polyclonal and each of the clones represented within these infections occur at different proportions. With the aid of genetic markers such as single nucleotide polymorphisms (SNPs) or microsatellites, parasite clonal genotypes can be identified. In this study, the genetic markers of choice are SNPs. Using a method that can quantify these SNPs representing the different clones occurring at different proportions in an isolate, then each of the clonal genotypes can be determined. Microsatellites were also used as additional markers in the study.

In this thesis, 1.Genetic markers (SNPs) across the *P. falciparum* genome were identified (Chapter 3); 2. Pyrosequencing™ was validated as a technique that would enable the identification of each genetically distinct clone represented in an infection by assigning proportions to the SNPs representing each genetically distinct clone and enabling the identification of parasite clonal genotypes in every isolate analysed. This was validated using laboratory prepared clone mixtures of *P. falciparum*. In addition; the progeny from a cross derived from genetically characterised 3D7 and HB3 isolates was analysed in preparation for the analysis of the field isolates (Chapter 4). , 4. In Chapter 5, field isolates were tested and clonal genotypes identified using both SNPs and microsatellites. A detailed population genetic analysis was also performed and finally in Chapter 6, evidence for correlation between the genetic relationships of these parasites and geographic distance was investigated.

The results from field isolates summarised in this thesis were from analysis of 54 isolates; 7 samples collected from Cameroon, 13 from Kenya and 34 from Mali. The data consists of 13 SNPs analysed by Pyrosequencing™ and 8 microsatellites. 84 clonal genotypes were identified by both genetic markers from the analysed isolates.

Analysis of both SNPs and microsatellites revealed that microsatellites were more informative than the SNPs based on the observed allelic richness and heterozygosity (genetic diversity) across all loci analysed. The overall F_{ST} value was 0.061 using SNPs and 0.043 by microsatellites analysis. These values were low but consistent with what is typically observed in African *P falciparum* populations.

Finally, analyses of the combined data set revealed that no statistically significant levels of spatial autocorrelation existed within the studied parasite populations. However, there was evidence of within host mixed parasite infections exhibiting a high level of genetic relatedness compared to between host infecting clones.

I keep the subject of my inquiry constantly before me, and wait till the first dawning opens gradually, by little and little, into a full and clear light.

- ***Isaac Newton***

1. INTRODUCTION

1.1. Chapter synopsis

The chapter begins with a brief background on discovery and history of malaria, the debilitating effects of the disease, the different types of *Plasmodium*, and, the lifecycle of the parasite in the female *Anopheles* mosquito and in the human host.

Next, important principles of malaria transmission developed using epidemiological concepts and mathematical modelling will be discussed including their implications in malaria control programs. Further the spatial distribution of malaria parasites revealing the focal and heterogeneous aspects of the disease and how obtaining this information can improve malaria intervention programs will be outlined.

Since the objective of this thesis is to use the genetic relationships of malaria parasites to define dispersal of these parasites in different geographic settings, the multiplicity of infection and how this confounds detection of genetically distinct clones will be discussed. In addition, a review is provided detailing advances in technology for genome-wide exploration and subsequent ease in detection of parasite clones from multiclonal infections.

Finally, the project concept which provided the impetus for this study and, the intended objectives in this thesis are given.

1.2. Background

Charles Louis Alphonse Laveran, a French army doctor, identified malaria parasites in fresh blood in 1880. He must have had remarkable eyesight as this was 10 years before Dimitri Romanowski developed the methylene blue staining method. It was not until 17 years later in 1897 that Ronald Ross demonstrated that these parasites were transmitted by mosquitoes providing insight on the lifecycle of these parasites then known as *Plasmodium* (historical perspective obtained from Bruce-Chwatt LJ, 1981; Capanna, 2006).

130 years since Laveran's discovery, malaria still remains a global burden with statistics according to the recent WHO report 2010 estimating that it affects approximately 300 million individuals and causes approximately 700,000 deaths annually. Its greatest impact is exerted in sub-Saharan Africa with most cases reported in children less than five years of age and pregnant women. The disease is caused by protozoa of the genus *Plasmodium* and is transmitted by female *Anopheles* mosquitoes.

Plasmodium falciparum is of great significance as it is highly pathogenic causing the most severe form of malaria with the highest number of reported cases predominantly in Africa. The malaria disease caused by this parasite is characterized by cyclic patterns of fevers and chills, without a clear remission between peaks of fever. *P. falciparum* malaria is highly prevalent in Africa due to the fact that *Anopheles gambiae*, the most important vector in Africa, has a strong preference to blood feeding on humans rather than other animals, with a penchant for feeding and resting indoors increasing its chances of human interaction. It also has a long

lifespan, increasing such a mosquito's chance of becoming infected with human malaria parasites and of transmitting it to another human being (Beier, 1998). In addition, Africa has temperate and stable weather conditions enabling the *Anopheles* host to thrive and hence transmit the parasites to the human host.

The other species of human malaria parasites are *Plasmodium vivax, Plasmodium ovale, Plasmodium. malariae* and *Plasmodium. knowlesi* (Singh *et al*., 2004). *P. vivax* and *P. ovale* cause periodic fevers recurring every 48 hours but *P. malariae* fevers recur every 72 hours. Recent findings indicate that *P knowlesi* is zoonotic being transmitted from macaque monkeys to humans by the vector *Anopheles leucosphyrus* (Coatney *et al*., 1971; Cox-Singh & Singh, 2008a). This strain causes cyclic periodic fevers every 24 hours (Jongwutiwes *et al*., 2004; Cox-Singh *et al*., 2008b).

1.2.1. Lifecycle of malaria parasites

The *Plasmodium* life cycle begins with a bite from an infected female *Anopheles* mosquito to the human host during a blood meal (**Figure 1.1**). About 10-100 sporozoites are injected and deposited into the skin of the human host (Doolan *et al*., 2009; Nkhoma *et al*., 2012), locate a blood vessel and migrate to the liver. This initiates asexual reproduction (schizogony), an asymptomatic phase.

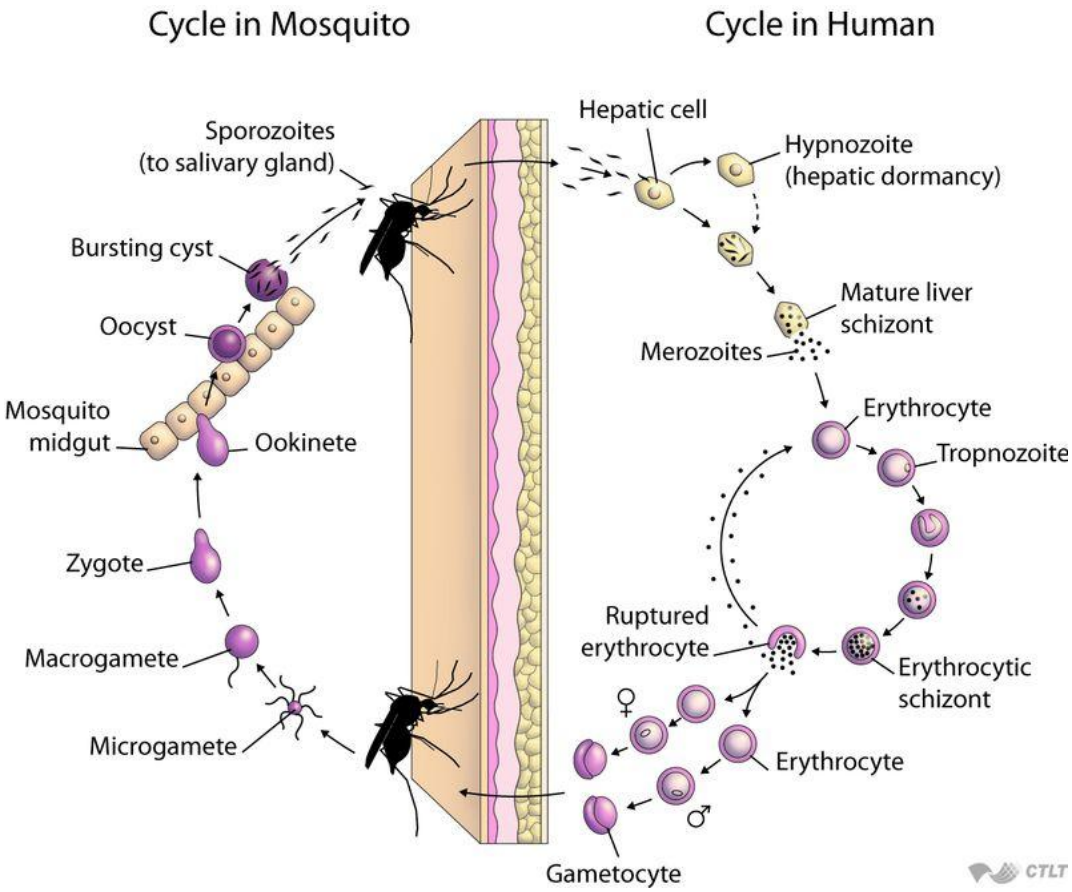


Figure 1.1 Lifecycle of *P. falciparum* showing the asexual phase in the human host and the sexual stage in the female *Anopheles* mosquito (Wirth 2002).

Each sporozoite invades a single hepatic cell developing and multiplying to 10,000-30,000 merozoites (reviewed by Jones & Good 2006). In *P. vivax* and *P. ovale*, some of the parasites can remain in the liver for years as hypnozoites and may cause recurrent episodes of malarial disease. The schizogony stage in the liver takes approximately 6 days in *P. falciparum*, 8 days in *P. vivax*, 13 days in *P. malariae* and 9 days in *P. ovale and P. knowlesi* respectively (Borrowed from the malaria site: www.malariasite.com).

The merozoites in the liver are released from ruptured hepatocytes into the blood stream and immediately invade red blood cells (RBCs) initiating the blood stage infection. The merozoites in the RBC develop to a 'ring' stage and then to a trophozoite which replicates to form a multinucleated schizont containing an average of 10 merozoites (review by Greenwood *et al*., 2008). When the schizont matures, the erythrocyte ruptures and releases merozoites that infect new erythrocytes. The characteristic periodic fevers that are the signature of malaria are precipitated by synchronous parasite development and erythrocyte rupture which releases new merozoites, malaria antigens and toxic metabolites.

Rather than maturing into schizonts, some of the trophozoites develop to male and female sexual gametocytes. The haploid gametocytes are ingested as part of the female mosquito's blood meal and initiate sexual reproduction (sporogony) in the mosquito's gut (**Cycle in mosquito, Figure 1.1**). The male gametocyte nucleus undergoes 3 rounds of DNA replication producing eight nuclei with each forming a flagellum in a process known as exflagellation. Thus, 8 motile male microgametes are produced. The female gametocyte matures to a macrogamete. The male and female gametes fuse to form a diploid zygote. At this stage, if a mosquito ingests two genetically distinct parasite clones, recombination takes place leading to emergence of novel genotypes. The motile zygote (ookinete) migrates through, and encysts on, the outer surface of the gut wall (as an oocyst). Asexual division within this oocyst produces numerous haploid sporozoites which migrate to the mosquito's salivary gland ready to be transmitted to the human host in a blood meal. The sporogonic phase lasts between 8-15 days in the mosquito (Carter & Graves, 1988; Eichner *et al*., 2001).

1.2.2. Pathophysiology of malaria

Malaria can be categorised as:

1. Uncomplicated malaria:

This presents in the form of fever which occurs following bursting (lysis) of schizonts in the blood (**Figure 1.1**) to release merozoites that invade RBCs. This lysis of the infected RBCs (iRBCs) is usually accompanied by release of toxins which initiate a paroxysmal response in the human host. These toxins have been observed to initiate a toxic shock response where monocytes activate the production of cytokines such as the pro-inflammatory TNF-α or activate the production of reactive nitrogen intermediates necessary for killing of asexual and sexual parasites with the exception of ring stages which are relatively heat-resistant (Karunaweera *et al*., 2003; reviewed in Langhorne *et al*., 2008). The fevers may also be accompanied by headaches, general malaise, body aches, sweating. Excessive hemolysis may lead to severe anaemia and enlargement of the spleen. If malaria is left untreated, this may lead to severe disease.

2. Severe and complicated malaria:

Two factors are associated with severe malaria:

a. Reduced deformability of iRBCs - An infected RBC undergoes changes as the parasite grows within it. In *P. falciparum* infections, iRBCs have been observed to become rigid and spherical losing their fluidity unlike the normal RBCs that are biconcave enabling them to squeeze through microvasculature (Paulitschke & Nash, 1993). This mechanism enhances splenic clearance of

IRBCs but on the other hand causes blockage in microvasculature due to their inflexibility which manifests as severe malaria (review by Suwanarusk *et al*., 2004).

b. In *P. falciparum* malaria, mature forms of the parasite are not usually observed by microscopy as they adhere to the endothelium of microvasculature, a phenomenon known as cytoadherence (Berendt *et al*., 1994). In addition, uninfected RBCs and platelets have been observed to stick to iRBCs also known as rosetting which together with cytoadherence are believed to be involved in development of severe malaria (Wahlgren *et al.*, 1998 review). Major organs such as the brain, lungs, liver and the placenta undergo obstruction due to the sequestration of non-infected and infected RBCs affecting their function (Miller *et al*., 1994). Pathological examinations of individuals who have succumbed to severe malaria reveal this sequestration in major organs suggesting the association of severe malaria and cytoadherence (Fried & Duffy, 1996; Newbold *et al*., 1997). In the brain this manifests as cerebral malaria while in pregnant mothers as placental malaria. Cytoadherence and rosetting have been associated with presence of the surface molecules *Plasmodium falciparum* erythrocyte membrane protein 1 (PfEMP 1) that render the iRBCs 'sticky' causing them to form rosettes with non-infected RBCs and also, enabling them to sequester to capillaries or tissues (reviewed by Pasternak &, Dzikowski, 2009)

Constant malaria attacks in the human host exert a social, psychological, physical and economic burden (Carter & Mendis, 2002). The mental and physical capacities are affected reducing individuals' viability. These malaria attacks also expose the

individual to opportunistic infections and since in such a state the immunity is lowered, death ensues.

1.2.3. Management of malaria

Various tools and strategies have been developed or are still in the process of development to counter the debilitating effects of malaria. These are covered in this section.

(a) Anti-malarial drugs

These drugs may be classified according to the parasite stages they act against and include: (i) the pre-erythrocytic forms of *Plasmodium* especially for eliminating hypnozoites associated with *P. vivax* and *P. ovale* that may remain dormant in the liver for extensive periods causing disease relapse. Primaquine is the only drug known to act against these stages rendering it effective in eliminating hypnozoites (Wells *et al*., 2010). Prophylactic drugs recommended for use by travellers (presumed non-immune) to malarious regions fall in these categories (Castelli *et al*., 2010). (ii) Blood stage parasites e.g. quinine, chloroquine, sulfadoxine, pyrimethamine, mefloquine, halofantrine etc. Emergence of multi-drug resistant parasites has led to the use of different drug combinations including the discovery of the antimalarial properties of Artemisinin. Unfortunately reports of parasites resistant to ACTs have begun to emerge (Dondorp, 2009; Noedl, 2008). (iii) The sexual forms of the parasites. Primaquine is effective in eliminating gametocytes of all *Plasmodium* (White, 2008). Chloroquine and quinine have this activity against *P. vivax and P. malariae* but not *P. falciparum*.

NB. Although Primaquine is effective against most of the parasite stages, its use is limited when administered to individuals with the Glucose-6-Phosphate Dehydrogenase (G6PD) deficiency trait as it has been observed to trigger haemolytic crisis in these individuals.

(b) Vector control strategies

These have been developed to reduce contact between the female *Anopheles* mosquito and humans and include (i) Insecticide treated nets (ITNs) infused with insecticide which repel or kill mosquitoes that come into contact with them. One type of net is a conventionally treated net impregnated with insecticide that lasts temporarily, necessitating constant re-treatment every three months or at least once a year to ensure continued insecticide effect. The other one is a factory made long lasting insecticide treated net (LLINs) incorporated with insecticide and can last for up to three to five years (WHO, 2007). ITNs have been shown to reduce the malaria cases by up to 50% in comparison with untreated nets (Clarke *et al*., 2000; Greenwood *et al*., 2008). ITNs have contributed to reductions of child mortality cases by about 44% (Fegan *et al*., 2007) and increase an in child survival by 27% in Tanzania (Schellenberg *et al*., 2001). (ii) Indoor residual spraying (IRS) which acts by repelling or killing mosquitoes that enter and/or rest on the areas that the insecticide has been applied. It is therefore considered to be most effective against endophilic vectors such as *Anopheles gambiae* (http://www.who.int/whopes/en/; Pluess *et al*., 2010 review). (iii) Larviciding (source reduction) to curtail the development of mosquito larvae to adulthood using larvivorous fish such as the minnow also known as mosquito fish (*Gambusia affinis*), or the common guppy

(*Poecilia reticulata*) or bacteria such as *Bacillus sphaericus* and *B. thuringiensis var israelensis* [National Vector Borne Disease Control Programme (NVBDCP)].

(c) Vaccines

No malaria vaccine exists at the moment although the target set by vaccine developers is to reduce severe disease and death with a vaccine having 80% efficacy lasting more than four years by 2025 (MVTR, 2006). The current potential malaria vaccines are those that: (i) Target pre-erythrocytic stages to avoid the clinical manifestations of malaria in the human host. A circumsporozoite surface protein (CSP) target antigen recombinant vaccine named RTS,S/AS01E is currently in the phase 3 trial and has been shown to offer 51% protection against all forms of clinical malaria episodes in a phase 2 trial (in children aged 5-17 months) conducted in Kilifi, Kenya (Olotu *et al*., 2011). The RTS,S/AS01E is the most promising vaccine candidate on trial with preliminary phase 3 trial results revealing concordant results as the phase 2 trial with 55% reduction of clinical malaria episodes and 35% reduction of severe malaria cases in children of two age categories: 6-12 weeks and 5-17 months (The RTS,S Clinical Trials Partnership, 2011; White, 2011). (ii) Targeting the erythrocytic stages for elimination of the asexual blood stages of malaria parasites thus reducing the clinical manifestation of the disease. Potential vaccine molecules include those involved in the merozoite invasion of RBCs e.g. Apical Membrane Antigen 1 (AMA1), Merozoite Surface Proteins 1 & 2 (MSP1 & 2) and Recombinant Rhoptry-associated Protein 2 (RAP2) (WHO, 2011). Others such as one form of the surface antigen PfEMP1 known as VAR2CSA which binds to the chondroitin surface antigen (CSA) in the placenta offer possible vaccine candidates for malaria protection of pregnant women (Duffy, 2007). Of great

interest is MSP1 which has been implicated in eliciting a strain specific protective immune (SSPI) response i.e. protection achieved from the exposure of one genetically distinct strain does not ensure protection from another genetically distinct strain. MSP1 has been associated with SSPI in studies conducted using animal models and the rodent malaria *P. chabaudi chabaudi* (Cheesman *et al*. 2009; 2010). Since MSP1 is highly polymorphic, the design of a hybrid vaccine encompassing all the variants present in natural *P. falciparum* infections has been proposed (Cowan *et al*., 2011). (iii) Targeting the sexual and mosquito stages essential for reducing the chances of gametes being fertilised thus reducing transmission of the parasites, which is why they are known as transmission blocking vaccines (TBVs). Surface antigens identified during the sexual stages of these parasites in the mosquito midgut are potential targets for these vaccines. The antigens, Pfs48/45 and Pfs230 expressed on the surface of gametocytes while still in the human host and in male and female gametes in the mosquito midgut constitute prospective transmission blocking vaccine candidates (reviewed by Carter, 2001). Antigens expressed on the surface of zygotes and mature ookinetes in the mosquito midgut following fertilization have also been identified namely Ps25 and Ps28 (reviewed by Carter, 2001). These antigens are favoured vaccine candidates as they have not been exposed to the human host immune system enabling them to develop evasion mechanisms. Animal vaccine trials using these antigens reveal strong transmission blocking activity (Miura *et al*., 2007) whereas in humans similar findings were observed but trials were suspended due to reported adverse reactions (Wu *et al*., 2008)

The current focus by the Malaria Eradication Research Agenda (malERA) is to eradicate the disease (Alonso *et al*., 2011). MalERA's goal is to develop a vaccine

targeting antigens in: (i) the sexual stages in the female *Anopheles* mosquito, (ii) the pre-erythrocytic stages with emphasis on *P. vivax* and *P. o*vale that are characterised by hypnozoites in the liver stages. VMP001/AS01B, a circumsporozoite based vaccine, is the only *P. vivax* vaccine under clinical trial at the moment (http://clinicaltrials.gov/ct2/show/NCT01157897), (iii) the erythrocytic stages reducing the clinical manifestations of the disease. This vaccine has been classified under the broad concept broad concept of Vaccines that Interrupt Malaria Transmission (VIMT) (The MalERA consultative group on vaccines, 2011).

1.3. Principles of malaria transmission

1.3.1. The basic malaria reproductive number (R_0)

The concept of R_0 initially developed by Ross (1911) was further expounded by Macdonald (1957). R_0 is defined as the number of new cases generated from an existing case of malaria in a naïve population. R_0 is determined by efficiency of a vector in transmitting malaria i.e. the vectorial capacity, and the interaction of the parasites and the human host, leading to the parasites infecting the vectors i.e. the human capacity of transmission. R_0 therefore defines the intrinsic capacity of an environment to support malaria transmission.

Malaria transmission is sustained when:

$R_0 > 1$

and, malaria transmission cannot be maintained when:

$R_0 < 1$.

Different places have different R_0 values and these can fall between 0 (a region with no intrinsic capacity to support malaria transmission e.g. with no mosquitoes), to possibly 100 to 1000.

R_0 helps establish the amount of work that needs to be done to reduce R_0 to < 1. It follows that places with an intrinsic capacity for malaria transmission (R_0) of less than 1 require no work at all in ensuring no malaria transmission. At the other end of the spectrum, places with R_0 of 100 to 1000 would require massive work to reduce it to less than 1 and ensure no malaria transmission.

The R_0 equation is given as:

$$\mathbf{R_0 = m.a^2.\ [p^n/-\ ln(p)].\ \pi.\delta.\varepsilon}$$

Where:

m = the number of female *Anopheles* mosquitoes' bites per person

a = number of humans that are bitten by a single female *Anopheles* mosquito in a day

p = the chance of the female *Anopheles* mosquito surviving for one day

n = how long in days it takes from a female *Anopheles* mosquito's blood meal containing gametocytes to the development of infective sporozoites in the mosquito's salivary glands (extrinsic cycle)

π = the probability of human susceptibility to infection (In the lifecycle (**Figure 1.1**), between a mosquito bite with sporozoite injection to the liver stage)

δ = how long in days potentially infectious gametocytes are in the blood of a malaria naïve person (This is equivalent to the reciprocal of the recovery rate, **1/r,** in Macdonald, 1957).

ε = the probability that the gametocytes in the human host will infect mosquitoes

The equation is composed of the vectorial capacity and the human capacity for malaria transmission.

1. The vectorial capacity

The vectorial capacity defines how good local mosquitoes are in transmitting malaria parasites to humans. From the R_0 equation above, the vectorial capacity is as shown below:

$$\mathbf{m.a^2.[p^n/-lnp]}$$

The reduction of vectorial capacity was the basis for the malaria eradication program recommended by WHO in the 1950s.

2. The human capacity of transmission

This is how effective humans are in maintaining the parasites and transmitting them to the vector. From the R_0 equation above, the human capacity of transmission is represented by:

$$\pi.\delta.\varepsilon$$

These are the targets of drugs and vaccines that reduce malaria transmission.

The parameters supporting the transmission of malaria have a multiplicative effect as shown in the equation. Therefore, using multiple malaria control methods would be highly effective as they would attack and reduce the different aspects necessary for sustaining malaria transmission leading to the reduction of $R_0 < 1$.

Sub-Saharan Africa is characterised by high malaria transmission intensity which offers a great challenge in the implementation of malaria control programs. Heightening this problem is poverty, bad governance, lack of proper health care services and lack of proper government structures. It is against this background that

proper malaria interventions require to be established and sustained to reduce the malaria burden.

Successful malaria control programs have been observed in the past to be evidence based. A good example is the malaria eradication campaign by William Gorgas during the construction of the Panama Canal (Gorgas, 1918). Identifying the vectors responsible for spreading malaria in an area, their feeding, resting and breeding patterns and their interaction with the human host is important prior to the implementation of malaria control programs.

1.3.2. Strategies ensuring effective malaria control

Malaria eradication efforts have shied away from regions of intense malaria transmission, like Sub-Sahara Africa (WHO, 2008). The strategies discussed below would ensure effective dissemination of the currently available malaria interventions.

1.3.2.1. Integrating malaria interventions

WHO has recommended integrating the previously discussed tools, hence ensuring an attack against all the parameters defined in the R_0 equation responsible for sustained malaria transmission. It is logical and sensible to scale-up and integrate the existing malaria interventions to ensure that in areas of high malaria transmission ($R_0>>1$), the value of R_0 is reduced and maintained at less than 1.

Although integrating malaria interventions is sensible, implementing it in areas of high malaria transmission intensity is costly. A study conducted in Aneityum Island,

a low and unstable malaria transmission area in the South Pacific, achieved success in eliminating malaria (Kaneko, 2010). However, this was feasible because the island has low malaria transmission intensity and only 718 inhabitants, making it possible to carry out effective malaria intervention through constant monitoring and surveillance. A study conducted in Bioko Island, Equatorial Guinea, an area of high malaria transmission, successfully achieved reduction over 2 to < 5 years in child mortality rates, mosquito abundance and infection prevalence (Kleinschmidt *et al*., 2009). In spite of the 4 year intensive effort in this project, which garnered support from the local government as well as from internal and external donors, malaria is yet to be eliminated from this island (Steketee, 2009). The project can indeed be considered a success but achieving this in areas of stable malaria transmission such as sub-Saharan Africa requires massive scale-up. Reduction in malaria risk has also been observed in Zanzibar (Bhattarai *et al*., 2007), Gambia (Ceesay *et al*., 2008), Kilifi, Kenya (O'Meara *et al*., 2008), Rwanda (Sievers *et al*., 2008; Otten *et al*., 2009) and Ethiopia. These reductions have been precipitated by massive scale up of interventions fuelled by improved international donor support enabling availability of long lasting impregnated bed nets (LLIN), artemisinin-based combination therapy (ACT), & rapid diagnostic tests (RDT). Although the feasibility of achieving malaria elimination in high malaria transmission such as sub-Saharan Africa remains unknown, even so, effective malaria control in such areas is achievable as outlined in the next section.

1.3.2.2. Malaria interventions by targeted treatment

In this thesis, transmission of malaria will be shown to be focal (**Section 1.3.3**); for example individuals closest to mosquito breeding grounds are at high malaria risk. In addition, within these high malaria risk areas, malaria transmission is heterogeneous as malaria incidences have been observed to cluster in certain households and not others as shown in **Section 1.3.4**. In light of this discussion, applying malaria interventions by identifying and targeting these high malaria risk areas (malaria hotspots) is highly recommended (Woolhouse *et al*., 1997; Smith *et al*., 2005; Bousema *et al*., 2010; Bousema *et al*., 2012).

Integrating malaria interventions coupled with proper deployment of these interventions by targeting malaria hotspots would lead to marked reduction of malaria in Africa.

1.3.3. Focal transmission of malaria

One of the fundamental aspects of malaria transmission is that it is focal. The female *Anopheles* mosquito is haematophagous as it requires blood from a human host for egg formation, after which it identifies the nearest breeding site to lay its eggs (oviposition). It is during this time that it might encounter a human host carrying *Plasmodium* parasites which it ingests as it takes its blood meal. The interaction of the vector and the human host is therefore dependent on the presence of a suitable and productive breeding site. It is on this basis that malaria transmission is focal. It follows that, after oviposition, the vector secures its next blood meal from the human host residing within its limit of movement to and from the breeding site.

A focus of malaria transmission is therefore a geographic region with the capacity to sustain malaria transmission as it has suitable climatic conditions to support vectors, suitable breeding grounds and human habitation. Such areas have a basic reproductive number (R_0) >1 (discussed in **Section 1.3.1**).

Malaria risk tends to decline with distance from a breeding site because it is dependent on mosquito density and the ratio of female vector *Anopheles* to the number of people at a particular location, as represented by m in the equation for R_0 (**Section 1.3.1**).

$$m = \frac{\text{No. of Anopheles mosquitoes}}{\text{No. of people}}$$

Thus individuals residing in an area characterised by high mosquito abundance (high m value) are at high malaria risk compared to individuals in an area having fewer mosquitoes (low m value).

In Africa, *Anopheles* mosquitoes have been observed to travel ~1km from a breeding site in search of a blood meal (Trape *et al*., 1992; Takken *et al*., 1998). This may be attributed to the behavioural patterns of *An. gambiae*, the predominant vector in Africa, which preferentially breeds in transient sunlit water collections that are widespread and close to human habitation (Minakawa *et al*., 1999); Mutuku *et al.*, 2006).

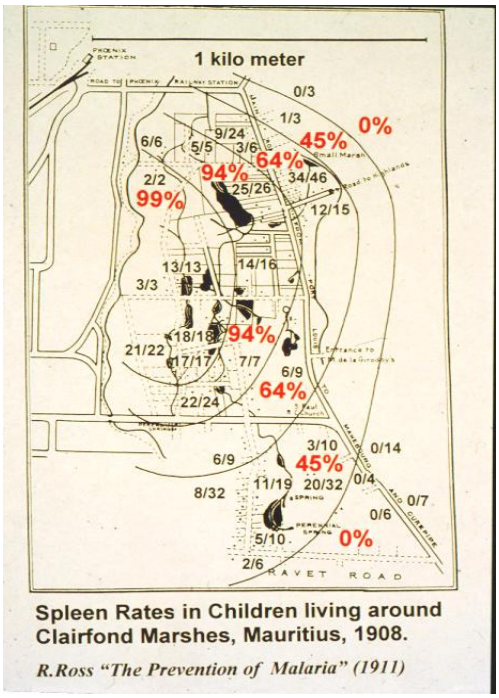


Figure 1.2 Malaria prevalence reported as percentage spleen rates in children. This figure shows a focus of malaria transmission where the prevalence of malaria reduces with increasing distance from the breeding sites (darkened areas in black). In the locality given above, malaria is being transmitted to a distance of approximately 1km (Ross, 1911).

To reiterate, the observation made by Ross, 1911 (**Figure 1.2**) indicates that a focus of malaria transmission is within a radius of ~1km from a breeding site but beyond its border malaria transmission is not sustained i.e. 0% malaria risk with $R_0 < 1$ values. Conversely, in regions where productive breeding grounds are isolated and located a great distance from human habitation such that mosquitoes have to travel far in search of a blood meal, defining a focus of malaria transmission is difficult. In this case, mosquitoes are widely dispersed and foci of malaria transmission cover a greater distance and may overlap (Carter *et al*. 2000). For instance, individuals at highest malaria risk were found to be located at a distance of 2km from a breeding site in Gambia (Thomas & Lindsay, 2000), 3km away from breeding grounds

(microdams built as sources of water) in Ethiopia (Ghebreyesus *et al*., 1999) and 5km away in South America (Charlwood & Alecrim, 1989).

1.3.4. Malaria heterogeneity within foci of transmission

Malaria has also been observed to be clustered within certain households and not others (**Figures 1.3**). The degree of contact between human hosts and the mosquito varies with some mosquitoes being attracted to certain houses. Poorly constructed houses that are not properly sealed allowing entry of the vectors have been observed to report with higher malaria cases (Gamage-Mendis *et al*., 1991).

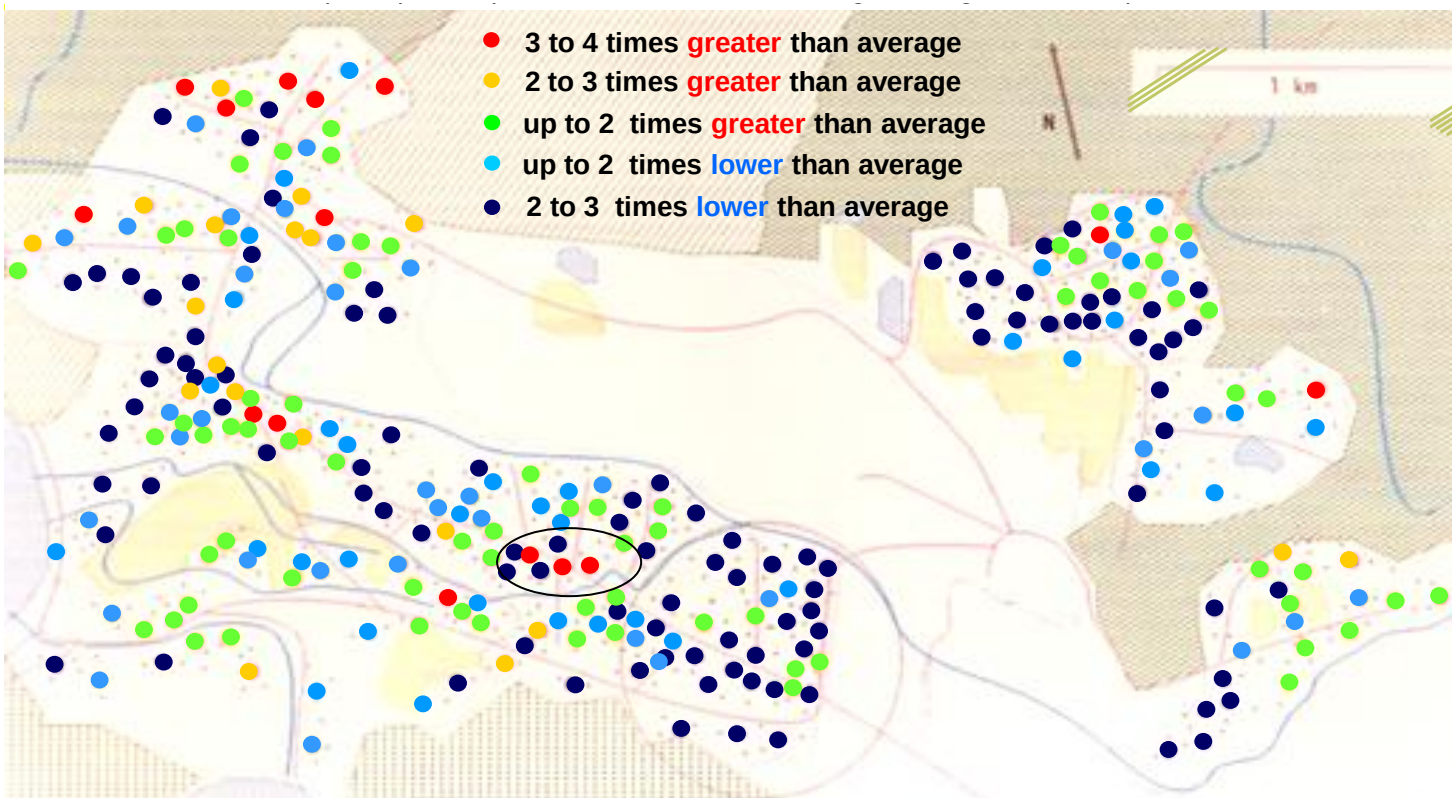


Figure 1.3 Study carried out in Sri Lanka showing heterogeneity of malaria transmission within a 1 km region. Each coloured point represents a total of 10 individuals in neighbouring houses. Red points represent the highest number of reported malaria cases while the black points represent the lowest as described above. Note the black circled region above. Households located within close distance of each other present with varying degrees of malaria reported cases ranging from 3-4 times higher than average (red dots) to lower than average cases (black dots). Malaria cases are therefore clustered within certain households and not others. This is an indication of heterogeneity within a focus of malaria transmission (Figure adapted from Gamage-Mendis *et al*., 1991).

Certain individuals have also been observed to attract mosquitoes because of their body odours (Takken & Knols, 1999; Mukabana *et al*., 2002). Mosquitoes have also been observed to preferentially feed on individuals harbouring gametocytes (Lacroix *et al*., 2005). Some studies have also suggested that people with certain blood groups are more prone to mosquito bites than others (cited by Mukabana *et al*., 2002). RBC polymorphisms that confer protection against malaria to certain individuals e.g. the sickle cell trait have also been observed to be responsible for the uneven distribution of malaria cases within foci of transmission (Kreuels *et al*., 2008). Other factors that contribute to heterogeneity include use of mosquito repellents or bed nets (Srinivas *et al*., 2005) and even pregnancy (Ansell *et al*., 2002).

Evidence of this clustering of malaria cases is given in different studies conducted in Africa and Sri Lanka (Gamage-Mendis *et al*., 1991; Lindsay *et al*., 1995; Gunawardena *et al*., 1998; Ghebreyesus *et al*., 2000). A study conducted in a village at Dielmo, Senegal revealed that children between the ages of 0-2 years had differing malaria episodes. Most children had 1-2 malaria attacks during the study period but two children reported with 20 and 40 clinical episodes respectively (Trape *et al*., 2002). In a study conducted in Kenya over a period of 3-5 years within a small geographical region (~2km), some children did not have any malaria episode whereas approximately 21% of children in the same age group experienced >2 episodes of malaria over different seasons and were termed the 'more susceptible' group (Mwangi *et al*., 2008). The more susceptible children were likely to be highly exposed to malaria in comparison to the children who were negative during the study

period and this evidence of spatial clustering was observed throughout the study period (Mwangi *et al*., 2008)

The heterogeneity in transmission of infectious agents occurs in other host-vector interactions such as humans, black flies and onchocerciasis (Bockarie *et al*., 1990); cattle, tsetse flies and trypanasomiasis (Wacher *et al*., 1994); humans, snails and schistosomiasis (Wilkins *et al*., 1987; Chandiwana & Woolhouse, 1991). This variation in transmission of pathogens was quantified from observations made on host-vector contact. Mathematical simulations were made from the data revealing the consequences of heterogeneous host-vector interactions. The results showed that typically 20% of the human host population has a net potential to transmit approximately 80% of the infectious agent (Woolhouse *et al*., 1997). Further studies conducted on *P. falciparum* malaria transmission (Smith *et al*., 2005; Smith *et al*., 2007) corroborate with the findings by Woolhouse *et al*., (1997) in Tanzania.

So far, evidence for heterogeneity in malaria transmission has been shown. Where only a few individuals may harbour malaria infections, they play a major role as they serve as reservoirs of infection in field settings.

1.4. The spatial distribution of Plasmodium

In the previous section, the principles of malaria transmission were discussed. Since malaria transmission is non-uniform, prior to carrying out any intervention, it is important to map out areas at great malaria risk. This way, malaria control efforts can be tailored to the needs of local settings to maximise their effectiveness.

Mapping malaria risk in Africa (MARAIARMA URL; http://www.mara.org.za) and Scientists for Health and Research for Development (SHARED URL; http://www.shared.de) are databases established to reveal the spatial heterogeneity of *P. falciparum* across Africa (le Sueur *et al*., 1997). The maps were generated from a meta-analysis conducted on available annual entomological inoculation rates (AEIR) data as measures of malaria transmission intensities across Africa. The AEIR is calculated as previously described elsewhere (Drakeley *et al*., 2003; Shaukat *et al*., 2010):

EIR = mas **(refer to R_0 equation in Section 1.3.1)**

m = Number of *Anopheles* per person (density of human-biting *Anopheles*)

a = Mean number of persons bitten by one female *Anopheles* per day (human blood index

N.B. **ma** = human biting rate

s = Proportion of sporozoite infected *Anopheles* (Sporozoite rate)

NB. To obtain AEIR the above parameters are averaged over a year

The malaria risk maps were generated with the aid of a Geographical Information System (GIS), a computer based programme that allows spatial data manipulation, interrogation, assemblage and visualisation (http://www.esri.com/).

These maps were of great benefit to guiding malaria control efforts as they revealed the population at risk of malaria in sub-Saharan Africa. This data highlighted the relevance of incorporating the spatial aspects of malaria distribution for the targeting of malaria interventions (Hay *et al*. 2000a, Hay *et al*., 2000b). In addition, the temporal aspect of malaria distribution was included in the map data, which recognised the changing patterns of malaria transmission during different seasons (Craig *et al*., 1999). For example, in the dry season, vector abundance is reduced

which leads to low malaria transmission. Conversely, following the rainy season collection of water puddles constituting suitable vector breeding grounds leads to elevated mosquito numbers and enhanced malaria transmission.

The MARA initiative led to the establishment of the Malaria Atlas Project (MAP), which contains global malaria risk maps (Hay & Snow, 2006) provided in the database: http://www.map.ox.ac.uk/. In addition to AEIR data, these maps were refined to include the parameter parasite ratio (PR), which is defined as the proportion of individuals confirmed malaria positive by the examination of asexual blood stages (rings) of malaria parasites with the aid of a microscope (Gilles, 1993). This parameter was used to provide a more specific definition of malaria endemicity (Hay & Snow, 2006) as a global baseline of malaria risk.

The determination of malaria distribution within small spatial settings has also been investigated. A retrospective study conducted in Gambia utilised data from previous surveys that identified potential mosquito larval habitats. Using this information together with GPS, a spatial autocorrelation extrapolation of EIRs and malaria risk over a 2km radii from these potential breeding sites was carried out (Thomas & Lindsay, 2000). In this study, high malaria transmission regions (closest to breeding sites) reported with low malaria clinical cases while low malaria transmission regions had more clinical cases. This is in contrast with what has been typically observed in Africa; i.e. individuals at greatest malaria risk are found in high malaria transmission regions. The reason given for this observation was that reduced exposure to malaria parasites in individuals residing in low malaria transmission regions caused a delayed development of acquired immunity to the disease (Thomas & Lindsay, 2000). A study conducted in Kilifi, Kenya based on data obtained from longitudinal studies

conducted over a 12 year period identified high malaria risk areas (Bejon *et al*., 2010). In this study, individuals reporting with more febrile episodes aggregated within regional clusters of about 1.3km.

A study conducted in Korogwe, Tanzania (Bousema *et al*., 2010) used GPS/GIS to determine expected clusters of malaria exposure within a radius of 2km (predicted flight distance of the *Anopheles*). Blood isolates collected from infants 8-16 weeks old were tested by enzyme-linked immunosorbent assay (ELISA) for IgG antibody levels against the antigens AMA-1 and MSP-1. The results indicated that clusters of high mosquito exposure within the 2km region corresponded to the highest sero-conversion rates (rate at which individuals develop antibodies against parasite antigens per year) and high malaria incidences (determined by AEIR). This is in line with what was observed in Kilifi, Kenya, where high mosquito exposure is associated with high malaria risk but is in contrast with the study by Thomas & Lindsay, (2000) in Gambia.

The above mentioned studies highlight the need to investigate the transmission of malaria within small spatial settings. Because of the heterogeneity of malaria transmission, different localities present with different transmission patterns. Prior to dissemination of malaria interventions, the local patterns of malaria transmission need to be established.

The prevalence of malaria parasites in different geographic settings in these studies were estimated by spleen rates (Ross, 1911), parasite ratios by microscopy (Gilles, 1993; Hay & Snow, 2006), AEIRs (Hay *et al*., 2000 a & b; Thomas & Lindsay, 2000) and using serological markers (Bousema *et al*., 2010). These strategies have

provided important information of the dispersal of malaria parasites over large geographic areas. However, since malaria parasites exist as mixed clone infections genotyping methods have also been developed to enable the detection and discrimination of parasite clones present per individual blood sample (isolate). This has facilitated the study of genetic variation in natural populations of malaria parasites at different geographic settings. In this thesis, the dispersal of malaria parasites in small spatial settings will be estimated by genotyping methods. These are favoured as genotyping platforms are fast, require small blood samples and are cheap. The next section reviews the generation of genetic variation in malaria parasites and implications for genotyping.

1.4.1. The genetic events of *Plasmodium* during the sexual stage of infection

During a blood meal, the female *Anopheles* mosquito ingests one or several genetically distinct clones from the human host. The only diploid phase in the life cycle is the zygote in the mosquito midgut. To retain haploidy, meiosis occurs shortly after zygote formation, resulting in sporozoites (refer to **Figure 1.1**).

Gametes arising from identical clones undergo self-fertilization to produce homozygous zygotes, containing identical alleles at all loci. In this case, recombination during meiosis does not have major genetic consequences, as the resulting haploid products possess genotypes that are indistinguishable from the parental clones (**Figure 1.4**). However, if gametocytes of two or more genetically distinct clones are taken up into the mosquito, genetic recombination as well as self-fertilization events among the gametes are expected. Genetic recombination during

meiosis of these forms will result in parasite progeny with novel gene combinations among the haploid progeny (**Figure 1.4**).

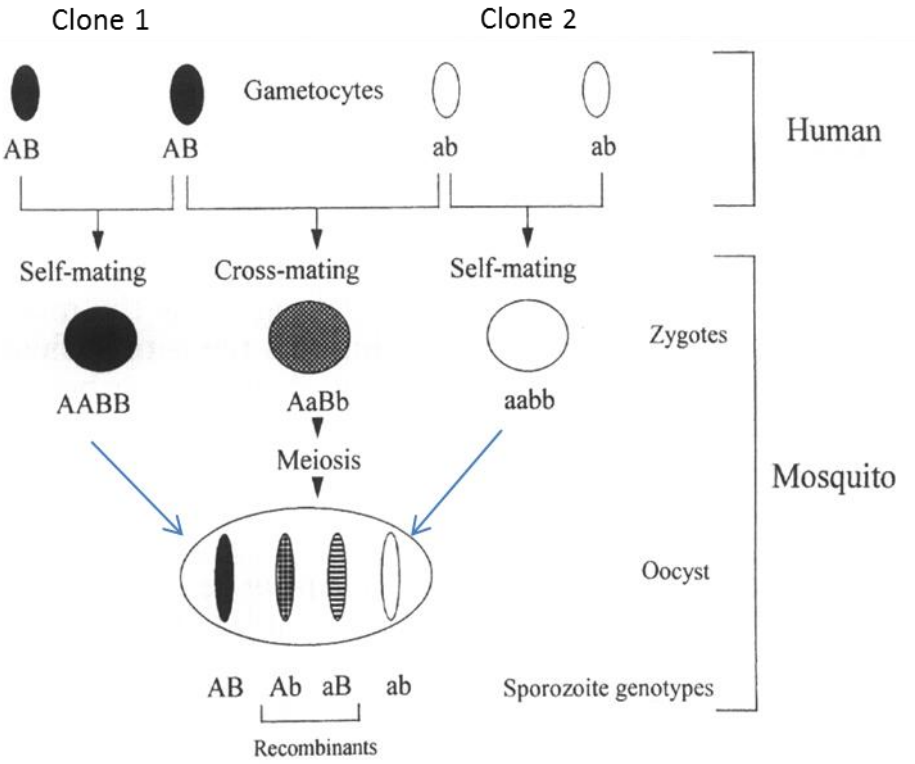


Figure 1.4 The process of genetic recombination in malaria parasites when 2 clones undergo cross mating in mosquitoes. A and a are alleles of one gene, and B and b are alleles of a second gene. One parent clone has genotype AB and the other ab; both produce gametocytes capable of infecting mosquitoes, between which mating can occur. Recombinants Ab and aB are produced following meiosis of diploid zygotes AaBb (Figure adapted from Babiker *et al*., 1999).

As shown in **Figure 1.4**, sexual recombination occurs in *Plasmodium* during meiosis and is important for the generation of novel genetic combinations. This occurs either by: (i) the independent assortment of unlinked genes on different chromosomes, (ii) crossing over events between linked genes on homologous chromosomes and (iii) recombination events within genes (intragenic recombination) leading to the generation of new allelic combinations (Walliker *et al.*, 1998).

The rate of genetic recombination in *P. falciparum* is high and has been estimated to be ~17kb per cM from an experiment conducted using the cross of the genetically distinct clones Dd2 & HB3 (Su *et al*., 1999). It has recently been estimated to be ~12.8 kb/cM using the progeny of the genetic cross between 7G8 and GB4 (Jiang *et al*., 2011). This means that there is a 1% chance of genetic recombination occurring every 12.8-17kb in the genome. Genetic recombination plays a key role in generating parasite variants that have major implications on the multiplicity of infection (MOI) in *P. falciparum* as discussed below.

1.4.2. The multiplicity of infection (MOI) in *Plasmodium*

The probability of parasites with novel clonal genotypes being generated in natural malaria infections depends on how often the mosquito takes in a blood meal containing more than one genetically distinct clone and the number of clones present in the human host at the time of the blood meal (Hill & Babiker, 1995). The question then arises as to how often multiple clonal genotype infections occur in nature. This is represented under different transmission conditions (inoculation rates) in **Table 1.1**.

Table 1.1 Frequency of multiclonal infections of *P. falciparum* in areas of different malaria transmission intensity

Parasite species		Entomological inoculation rate (EIR) #	Mean no. of clones per positive sample*	Proportion of positive samples with multiclonal infections	References γ
Plasmodium falciparum	Nigeria	>100	3.9 (1-9)	0.94	6
	Papua New Guinea	>100	3.9 (1-9)	0.8	11
	Senegal (Dielmo)	100-200	3.7 (1-8)	0.92	3,4
	Ghana	300	3.4 (1-8)	0.84	10
	Tanzania	>500	3.3 (1-6)	0.85	2
	Sao Tome	-	2.4 (1-6)	0.7	7
	The Gambia	>40	2.3 (1-4)	0.5-0.6	1
	Kenya	>30	2.1 (1-4)	0.7	8,9
	Senegal (Ndiop)	17	1.6 (1-4)	0.47	5
	Sudan	0.6	1.3 (1-3)	0.2	2
Plasmodium vivax	Papua New Guinea	>100	3.8 (1-8)	0.65	12
	Thailand	0.1-5	1.2-1.3 (1-3)	0.21	13
	China	0.1-1	1	0.05-0.06	13
Plamodium malariae	Malawi	1-5	2.1 (1-4)	0.71	14
	Thailand	0.1-5	1.2 (1-2)	0.19	14
	Gambia	1-2	1.2 (1-2)	0.2	14

* Mean (range in parentheses)

EIR values obtained from http://www.map.ox.ac.uk/

γ References: 1. Conway *et al*., (1991); 2. Babiker *et al.* (1999); 3. Ntoumi *et al*. (1995); 4. Trape *et al*. (1994); 5. Konate *et al*. (1999); 6. Engelbrecht *et al*. (2000); 7. Muller *et al*. (2001); 8. Kyes *et al*. (1997); 9. Mbogo *et al*. (1995); 10. Owusu-Agyei *et al.,* (2002); 11. Bruce *et al*., 2000a & b; 12. Kolakovich *et al*., 1996; 13. Yang *et al*., 2006; 14. Bruce *et al*., 2007.

The results depicted in **Table 1.1** indicate that there is a strong correlation between high malaria transmission intensity and the number of clones infecting individuals (MOI). This is as discussed below.

1.4.2.1. MOI in high malaria transmission areas

Individuals residing in regions of intense malaria transmission (EIRs > 100) are highly exposed to malaria parasites throughout the year. As shown in **Table 1.1**, individuals residing in these regions have an average of 3-4 parasite clones. A parasite infected blood isolate collected from a single individual can have as few as 1-2 genetically distinct parasite clones to about 8-9 clones (**Table 1.1**) or even more (upto 16 clones) as observed by Juliano and colleagues (2010a). Multi-clonal infections may arise from numerous mosquito bites with each bite injecting genetically distinct clones also known as 'superinfection' (Hill & Babiker, 1995; Nkhoma *et al*., 2012). Alternatively, a single mosquito bite may also contain numerous genetically distinct clones (Druilhe *et al*., 1998). Nevertheless, high parasitaemia levels are usually observed in these regions and high MOI values are expected because the higher the parasitaemia, the higher the chance of detecting multiple parasite clones (Contamin *et al*., 1995; Färnert *et al*., 2001).

The high parasite inoculation rates experienced in these regions lead to a pattern of development of clinical disease stratified by age. Children younger than five years bear the main burden of morbidity and mortality while older children and adults may develop partial immunity (Doolan *et al*., 2009). Protective immunity acts against blood stage malaria parasites and their products, reducing morbidity and mortality. Persistent inoculation of a human host with the parasites determines the rate of development of immunity (Baird *et al*., 1991). Immunity has also been observed to be both parasite species specific and 'strain' specific (Mendis *et al*., 1991). For every antigenically distinct strain, the human host must mount a specific immune

response. Further to this, for the immunity to be maintained, persistent parasite inoculation needs to be maintained.

Host immunity has an effect on the number of infecting parasite clones. Young children (under 5 years of age) whose immunity is not yet developed experience the highest clinical malaria episodes with some studies showing that they have low MOI compared to older children and adults (asymptomatic) (Magesa *et al*., 2002; Färnert *et al*., 2009). It has been suggested that pre-existing clones are usually suppressed or eliminated by the pyrogenic and toxic effects experienced during a clinical attack such that only a few genetically distinct clones are present (Felger *et al*., 1999; Arnot, 2002). On the other hand, older children and adults have been observed to harbour a high number of parasite clones as they experience less clinical episodes with the existing immunity sustaining the infecting clones at low proportions (Magesa *et al*., 2002; Färnert *et al*., 2009).

1.4.2.2. MOI in Moderate to low malaria transmission regions

Regions of moderate to low malaria transmission are characterised by EIRs of <50 to <1 and might occasionally experience peak malaria transmission during particular seasons e.g. during and after the rains. The average number of clones detected within individuals fall within the range of 2-3. Individuals residing in these regions have been observed to develop malaria immunity at a slower rate or not at all (Doolan *et al*., 2009). The number of parasite strains present in the population is therefore maintained at a low proportion. In addition, since immunity is slow to

develop, each infective bite to the human host may result in a malaria episode. Higher MOI values are usually observed during peak transmission periods (Vafa *et al*., 2008).

1.4.2.3. Other confounding factors of MOI

Human host polymorphisms that have arisen as a protective mechanism against severe malaria disease have also been implicated in the observed MOI in individuals. These include haemoglobinopathies such as the sickle cell trait (HbAS), α & β thalassemia or the enzymopathy, G6PD deficiency. Individuals having these polymorphisms rarely develop clinical malaria episodes and are therefore asymptomatic (Williams, 2006). Studies conducted to determine MOI in individuals possessing these erythrocyte variations have been contradictory. Ntoumi *et al*., (1997) in a study carried out in Gabon observed high MOI in HbAS individuals but in Senegal no such association was detected (Konate *et al*., 1999). So far no studies have been conducted to clearly affirm the assertion that the malaria associated human polymorphisms reduce MOI levels in individuals. However, since humans bearing these traits are usually asymptomatic and rarely present with clinical malaria episodes, they harbour malaria parasites and act as reservoirs of infection as discussed by Gouagna *et al*., (2010).

The number of clones present in an individual varies within the course of a clinical infection. For example in low transmission areas, the same clonal genotypes have been detected over long periods of time: 5-7 months (Babiker *et al*., 1998) to about 9

months (272 days) (Hamad *et al*., 2000) whereas in high malaria transmission areas clones were observed to persist for a single day to about 50 days (Daubersies *et al*., 1996), indicative of high recombination rate generating novel clonal genotypes. In asymptomatic individuals sampled over consecutive days of infection, different clonal genotypes were detected (Farnert *et al*., 1997; Bruce *et al*., 2000a; Magesa *et al*., 2002). This could be due to sequestration of the parasite, clearance of some parasites by host immune responses (Färnert *et al*., 2008) or in the case of *P. vivax* and *P. ovale* presence of hypnozoites in the liver that appear in the bloodstream after a prolonged period from time of initial parasite inoculation. Interestingly, similar parasite clones at the time of sampling were observed to be in synchrony i.e. either sequestered or in the blood circulation (Fänert *et al*., 1997; Bruce *et al*., 2000a). It is therefore recommended that repeated sampling of the same individual be carried out over several months in order to detect the repertoire of infecting clones present (Cole-Tobian *et al*., 2005). These studies give clear evidence that a single blood isolate collected at a single time point does not give the actual representation of the infecting parasite clones. This has implications especially in population genetic studies where small sample sizes representing the whole population are used to make inferences on, for example, the population structure of these parasites. Failure to have the correct sample representation may lead to poor interpretation and conclusions as far as these parameters are concerned.

Finally, the intake of antimalarial drugs affects MOI (Fänert *et al*., 2008). This has been demonstrated in a study carried out on individuals undergoing treatment for uncomplicated malaria in Uganda (Myrick *et al*., 2006). A striking feature in this study was the reduction of malaria parasites to levels below detection of microscopy

or PCR analysis during therapy. After several days following the clearance of the drug from the body (observed average 9.2 days), parasite recrudescence was observed i.e. the same parasite genotypes were observed before and after treatment (Myrick *et al*., 2006). This in turn affects the number of clones that can be detected with parasites occurring at very low proportions being missed out altogether yet they also contribute to the MOI levels.

1.4.3. Identification of *P. falciparum* clonal genotypes from mixed infections

The fact that malaria parasites exist as mixed clone infections necessitates the need to develop strategies that can identify parasite clonal genotypes present in every infection. This is important for the accurate representation of individual parasites when making inferences on parasite populations occurring in different geographic settings. The genetic composition of parasites in malaria infections can be investigated using genetic markers such as variation in the electrophoretic mobility of proteins e.g. enzymes, single nucleotide polymorphisms (SNPs) and microsatellites. These markers help identify parasites that are genetically related in a mixed clone infection by identifying parasites that carry the same gene allele combinations (genotypes). The detection of parasite clones present per blood isolate depends on (i) The ability of the markers used to discriminate the clones present (how polymorphic the markers are in elucidating clonal diversity) (ii) The ability to detect clones that are present at low proportions (minor clones).

Various molecular methods have been developed to identify these markers and have been used to quantify the proportions of genetically distinct *P. falciparum* parasites

occurring in mixed clone infections. As discussed below, these techniques have various detection limits.

1.4.3.1. Pre-genomic methods of detection

a) The electrophoretic forms of enzyme

Initial studies to determine MOI in *Plasmodium* investigated the variant forms of enzymes and antigens. Carter & McGregor in 1973 studied the electrophoretic forms of *P. falciparum* enzymes from field isolates collected in Gambia. Since blood form *P. falciparum* are haploid and the fact that these enzymes occur as single copy genes, each genetically distinct clone occurred as a single enzyme form (allele). Isolates containing more than one clone could be distinguished by the visualisation of the migratory patterns of these enzymes by starch gel electrophoresis. This study revealed that in the human host, these genetically distinct clones are present at different proportions.

Following the study by Carter & McGregor, (1973), other studies were conducted to determine the extent of *P. falciparum* genetic diversity in Africa (Carter & Voller, 1975; Sanderson *et al*., 1981; Creasey *et al*., 1990), South America (Creasey *et al*., 1990) and Asia (Thaithong *et al*., 1981; Creasey *et al*., 1990). Some of the conclusions drawn from these studies include that similar enzyme forms were detected across different geographic regions but at varying frequencies within the blood isolates studied. Secondly, some of the enzyme forms revealed extensive polymorphism compared to others which shows that some markers were more

informative than others. Indeed multiclonality was observed in these studies, with the study by Creasey *et al*., (1990) observing the highest MOI levels in Africa compared to South America and Asia. Carter & Voller, (1975) attempted to generate parasite clone genotypes from the observed enzyme combinations but this was only possible with two enzyme markers that occurred as two electrophoretic forms (a total of four possible enzyme clone combinations). The enzyme clone combinations obtained from Gambia (West Africa) and Tanzania (East Africa) differed in frequency within their countries of origin.

Although use of enzyme variants as genetic markers is cheap, easy and fast, few markers have been identified and can be visualised by electrophoresis (Shaw, 1970; de Meeûs *et al*., 2007). In *P. falciparum*, a total of six were used (Sanderson *et al*., 1981; Thaithong *et al*., 1981; Creasey *et al*., 1990). In addition, these markers display low polymorphism (Meeûs *et al*., 2007) limiting their ability to identify and distinguish the repertoire of parasite clones especially in areas of high malaria intensity. They also failed to reveal the degree of variation present between parasite clones as was observed by Creasey *et al*., 1990.

b) Determination of antigenic variation using monoclonal antibodies

Antigenic diversity was investigated using monoclonal antibodies (MAbs) to identify allelic serotypes of the *P. falciparum* surface antigens (MSP1, MSP2 & Exp-1) (Creasey *et al*., 1990; Conway & McBride, 1991). These antigens are known to have extensive allelic polymorphism such that the MAbs identify distinct antigen epitopes each representing a parasite clone. The MAbs were identified by staining using indirect immunofluorescence assay (IFA) (Mcbride *et al*., 1982). In the study by

Conway & Mcbride, (1991a & b) blood isolates obtained from an urban/periurban region in Gambia used MAb pairs to resolve the number of parasite clones present per blood isolates. 1-4 parasite clones could be identified.

Antigenic variants as markers of clonal diversity give more accurate results than enzyme forms due to the degree of polymorphism exhibited in these antigens such that the probability of encountering shared alleles per locus is greatly reduced (Conway & Mcbride, 1991b). The major setback with this approach is that it requires the laborious adoption of ring stage blood forms to culture for maturation to schizonts. This is because mature schizont forms were observed to react and stain intensely with IFA compared to the ring stage forms (Mcbride *et al*., 1982; Day *et al*., 1992). Another problem as highlighted in the study by Conway & Mcbride, (1991b) is the reduced ability to detect parasite clones at low parasite levels (1% of the total parasite number per isolate). However, this setback is encountered irrespective of technique used and advances in technology are underway to find a solution.

1.4.3.2. The genomic period

a. Polymerase chain reaction

The use of polymerase chain reaction (PCR) to genotype *Plasmodium* enabled the analysis of parasite DNA directly from blood samples without the need for DNA enrichment by culture establishment (Walliker, 1994). PCR technology has been utilised with the markers outlined below:

i. *P. falciparum* surface antigen genetic markers

The genetic markers encoding antigens on the surface of genes include: the most extensively studied: *P. falciparum* merozoite surface protein 1 (*PfMSP1*) & *PfMSP2* together with other genes such as: glutamate-rich protein (*PfGLURP*) & circumsporozoite protein (*PfCSP*) (Kimura *et al*., 1990; Marshall *et al*., 1994; Arnot *et al*., 1993). These markers exhibit sequence repeat length variation and can be distinguished by size using gel electrophoresis following PCR analysis. Allelic variation in the MSP1 gene is determined by the allelic families namely K1, MAD20 & RO33 and in MSP2 by the allelic types IC3D7 & FC27. These markers have been used for the determination of MOI mostly in Africa (Babiker *et al*., 1994, Creasey *et al*., 1990) and in South East Asia (Paul *et al*., 1995, Felger *et al*., 1994), see also **Table 1.1**. The findings revealed that there is a direct relationship between malaria transmission intensity, number of genetically distinct clones infecting individuals and, extensive clonal diversity observed between geographic regions at a continental level (Babiker & Walliker, 1997 and as outlined in **Table 1.1**). Further to this, in comparison to the previously used markers (enzyme forms and Mabs detection), by PCR, a higher number of genetically distinct parasite clones could be detected (see **Table 1.1**).

However, these genetic markers occur on the parasite surface antigens of the human blood stages and are therefore under high selection pressure to evade the host immune response (Hughes, 1992; Hughes & Hughes, 1995). They are therefore not ideal for studies of malaria parasites' population genetic structure. This is because the variation observed between the clonal genotypes could be due to frequency-

dependent selection maintained on these loci as a result of host immune response selection pressure obscuring gene flow patterns at a geographical level. In addition, a nested PCR is normally used for the amplification of these loci and in MSP1, in vitro recombination has been observed to occur in samples containing mixed genotypes (Tanabe *et al*., 2002). This may lead to an overestimation of the number of clonal genotypes identified and subsequently to the erroneous interpretation of parasite genetic diversity at a population level. Finally, routine PCR is not quantitative thus the DNA product does not give a representation of the proportions of parasite clonal genotypes per blood isolate.

ii. Microsatellites

Microsatellites are repetitive sequences across the genome of various organisms (Selkoe & Toonen, 2006). They comprise short tandem repeats of 2-6 nucleotides. They arise as a result of replication slippage where misalignment of the synthesized strand to the template strand leads to an insertion whereas a template strand – synthesized strand misalignment leads to a deletion (Levinson & Gutman, 1987). Microsatellites also arise due to the unequal crossing over events of homologous chromosomes during meiosis but this is common in alleles having large observable changes in their sizes compared to the small size changes between alleles arising from replication slippage (Jarne & Lagoda, 1996). They are favoured in population genetic studies as they have high mutation rates on the order of $10^{-3} – 10^{-4}$ per locus per gamete per generation (Avise, 2004), resulting in extensive allelic diversity that renders them highly informative. In *P. falciparum*, the microsatellite mutation rate

(μ) is higher than for SNPs and has been estimated to be $1.59x10^{-4}$ substitutions per microsatellite locus per generation (Su *et al*., 1999; Anderson *et al*., 2000a).

Microsatellites derived from loci in which polymorphisms appear to be neutral have been validated for use in *P. falciparum* population genetic studies (Anderson *et al*., 1999). Due to their highly polymorphic nature, they are appropriate for identifying closely related parasites (Su *et al*., 2007).

b. Sequencing technology

Advances in DNA sequencing and genotyping platforms have led to increased interest in use of SNPs as genetic markers (Vignal *et al*., 2002). The current technology is geared toward development of an accurate, high throughput, robust and less costly genotyping platform

A single nucleotide polymorphism (SNP) is a point mutation leading to the existence of an alternative form of a nucleotide at a genetic locus (unique position in a DNA sequence) between members of the same species or, in this context, between two genetically distinct *P. falciparum* clones. For genetic analysis, a SNP present in the population at a minor allele frequency of >1% is typically used (Vignal *et al*., 2002). In *P. falciparum*, comparison of 5 different parasite clones revealed a genome wide average of one SNP every 5.9kb (Mu *et al*., 2007). SNPs are highly conserved over many generations and in *P. falciparum*, the rate of mutation (μ) in SNPs has been estimated to be $1.2\text{-}1.7x10^{-9}$ substitutions per site per generation (Hughes & Vierra, 2001).

SNPs have recently gained favour in *P. falciparum* studies over microsatellites or the PCR analysis of antigenic genes due to their abundance enabling multilocus genotyping for the identification of parasite clones. Further, in population genetic studies, neutral markers are pertinent for the detection of population structure and gene flow patterns. The existence of numerous SNPs enhances the identification and analysis of neutral loci. The various technological advances for SNP identification and analysis in *Plasmodium* studies are discussed below:

i. The Sanger method of sequencing

The Sanger sequencing method also known as the 'dideoxy' or chain termination method (Sanger *et al*., 1977) formed the foundation of genome sequencing technology. This method involved a radioisotope (^{32}P) labelled primer complementary to the DNA being sequenced initiating the process. DNA polymerase sequentially added nucleotides to the growing strand until a terminator nucleotide was incorporated. Each of the reactions was performed in four different tubes containing the terminators A, C, G & T or ddNTPs. The numerous copies of DNA obtained were resolved by polyacrylamide gel electrophoresis and autoradiography (reviewed by Franҫa *et al*., 2002). This method was time consuming, expensive and laborious but major improvements were made such as addition of a PCR step for DNA amplification, radiolabelling the ddNTPs or fluorescently labelling the primers and replacement of polyacrylamide gels with capillaries housed in sequencing machines (also known as automated Sanger sequencing) (Hall, 2007).

The Sanger sequencing technology was used to analyse worldwide *P. falciparum* isolates (Mu *et al*., 2005). SNPs located in chromosome 3 of *P. falciparum* that had been previously identified from comparison of 5 worldwide isolates (Mu *et al*., 2002) were used in this study. In this study, a single clone was obtained and used for subsequent analysis. Haplotypes (allelic combinations) generated from the analysis of *P. falciparum* blood isolates collected from Africa, Southeast (SE) Asia, South & Central America and Papua New Guinea (PNG) revealed that these parasites possess unique haplotypes within their geographic regions indicative of population structure (Mu *et al*., 2005). These findings are similar to observations made by Anderson *et al*., (2000) using microsatellites. The Sanger sequencing technology although of great significance in genome wide studies is laborious, of limited throughput and time consuming necessitating the need to develop advanced sequencing technologies.

ii. Pyrosequencing

Pyrosequencing is a non-Sanger method for SNP detection within short sequence reads (about 20bp) developed by Ronaghi *et al*., 1996. Pyrosequencing was validated for genotyping and allele quantitation in mixtures of malaria parasites (Cheesman *et al*., 2007) and was used in this study. It facilitated the identification of *P. falciparum* mixed clone infections from 6 SNPs located in the $MSP1_{19}$ antigen in blood isolates collected from Mali (Takala *et al*., 2006). In most of the isolates, 1-2 haplotypes (1-2 parasite clones) were obtained. However, clear deciphering of individual haplotypes in isolates containing high MOI (>2 clones) was difficult (Takala *et al*., 2006). Further discussion on the use of this method in this study is in **Chapters 3 & 4**.

1.4.3.1. Advances in genomics – Next generation sequencing technologies

The Sanger method of sequencing is known as the first generation sequencing technology and its improved versions together with other recently developed sequencing technologies are known as next generation sequencing (NGS) technologies. The NGS strategies used so far in *P. falciparum* studies are reviewed below.

a. 454 Pyrosequencing technology

This is an advanced form of Pyrosequencing and is also known as the massively parallel pyrosequencing (MPP) (Zhou *et al*., 2010). It involves the fragmentation of DNA followed by the ligation of linkers at both ends of the fragments. The linkers are attached to a streptavidin bead and serve as primers for the PCR reaction. This is carried out in a picotiter plate where each cell contains a single bead (Hall, 2007). Following the PCR, each cell contains numerous copies of the DNA template. In the sequencing step, nucleotides are allowed to flow into the picotiter plate sequentially, whereby nucleotide incorporation leads to the production of light (photons) as occurs with the conventional Pyrosequencing strategy. This method is an improvement from Ronaghi's (1996) in that sequence reads of approximately 250bp can be obtained compared to 100bp. In addition, a large number of PCR amplified unique DNA templates are captured on beads (about 300,000 per picotiter plate) (Leamon *et al*., 2003) for subsequent sequencing in contrast to the single DNA sequence per bead obtained by the conventional Pyrosequencing method (Metzker, 2009).

The 454 sequencing technology was used for the massive parallel sequencing of *P. falciparum* MSP1 and MSP2 genes for the detection of infecting clones present in blood isolates collected from Malawi, a high malaria transmission region and Cambodia, a low malaria transmission region (Juliano *et al*., 2010). This method was compared to the conventional nested PCR approach for detection of allelic variation in these loci. There was a marked improvement in the number of parasite clones detected by the 454 sequencing method. In Malawi, there was a 5 fold increment in the number of clones detected by this method compared to the nested PCR method. In Cambodia, where parasite clones occurring at low proportions have a high probability of being undectected, there was a 6 fold icrease in the number of infecting clones detected per isolate (Juliano *et al*., 2010). Unlike the conventional Pyrosequencing technology, in this method, longer sequence reads can be obtained (upto 250bp). The sequences obtained can then be investigated for sequence diversity among the different observed parasite clones. This is a significant improvement from the nested PCR and fragment electrophoresis approach which has limited information as polymorphism is defined by size and not by exact sequences (Juliano *et al*., 2007; Liu *et al*., 2008).

b. Illumina technology

This technology is an advanced form of NGS currently in use at the Sanger institute (http://www.sanger.ac.uk/) and constitutes genotyping platforms such as Solexa. This involves the random cleaving of DNA into small fragments which are then ligated with adapters for attachment on a slide with flow cell channels. Amplification of each fragment generates clusters in each channel of the flow cell. The sequences are obtained by adding four fluorescently labelled reversible

terminators, DNA polymerase and sequencing primers. The clusters are then sequenced, with a camera capturing images of fluorescence emitted from each cluster. The obtained sequences are then aligned to a reference genome and assessed for quality. The setback with this technology is that the *P. falciparum* genome is AT rich such that the PCR is biased to amplifying regions with adequate GC content or regions with high sequence similarity misalign to the reference genome. Efforts have been made to eliminate the PCR step whereby the template DNA is attached to adapters that are partially double stranded together with the sequencing primer such that each strand is sequenced directly minus the cluster formation step (Kozarewa *et al*., 2009). This technology has recently been used by Manske *et al*., (2012) to develop a library of SNPs relevant for delineation of worldwide *P. falciparum* clones from field isolates. These SNPs are now available for population genetics studies in Africa as they are available in the open access website: http://www.malariagen.net/resource/10.

The Illumina GoldenGate technology has recently been used to detect and distinguish *P. falciparum* clones present in mixed infections (Campino *et al*., 2011; Nkhoma *et al*., 2012). The difference between this technology and Solexa sequencing is that instead of a solid flow cell, biotinylated DNA is immobilised on streptavidin beads like in Pyrosequencing™ technology then amplification and sequencing are performed in a 96 well plate.

1.5. The population genetics of P. falciparum

In this section, a brief review on the worldwide patterns of *P. falciparum* population structure is given. In addition, the basic concepts and underlying assumptions applied in population genetic studies that will be used in the analyses sections (**Chapters 5 & 6**) of this thesis will be introduced. Population genetic studies involve understanding how the distribution of genetic variation can explain interaction of individuals in space e.g. their mating systems or gene flow patterns.

One of the most extensive studies to define the population structure of malaria parasites was conducted using microsatellite loci that are assumed to be neutral (Anderson *et al*., 2000). This study was conducted in 3 African countries (known to have high malaria transmission intensity), 3 South American countries (moderate to low malaria transmission) and 3 sites from the Asian/Pacific region (one from Thailand (moderate malaria transmission) and two villages known to be malaria endemic in Papua New Guinea). The results showed that parasites clustered to their geographic region of origin at a continental level. Secondly, the degree of parasite genetic variation within Africa and Papua New Guinea was found to be high and little genetic differentiation was observed among localities in these regions. On the other hand, isolates obtained from localities in the Asian/Pacific region exhibited less genetic diversity and were highly genetically differentiated among the localities. Lastly, there was a correlation between the degree of genetic diversity and the transmission intensities of a region thus Africa and Papua New Guinea had the highest parasite genetic diversity, followed by South American and then Asia/Pacific

region. Mu *et al*., (2005) in another study using SNPs to determine *P. falciparum* population structure at a global scale mirrored the findings by Anderson *et al*., (2000).

Recently, Manske *et al*., (2012) also observed that at a global scale, malaria parasites clustered within their continents of origin. Also, parasite isolates obtained from within Africa were not genetically differentiated compared with those from Southeast Asia. Similar observations were made by Mobegi *et al*., (2012) in West Africa.

The above mentioned studies reveal that the level of parasite genetic variation across Africa is high. Since malaria transmission in Africa is heterogeneous, the genetic relationships of parasites in different localities are expected to vary. This is due to the focal nature of malaria transmission as discussed earlier. It can be assumed that each focus of malaria transmission represents a cluster of genetically interacting parasites that are highly related or, it comprises highly overlapping clusters whose boundaries cannot be clearly established. It is on this basis that this study was devised in an attempt to determine how malaria parasites are transmitted within different foci.

1.5.1. Population genetics concepts applied in this thesis

The interaction of malaria parasites in space is influenced by various underlying forces such as mutations, natural selection, genetic drift or gene flow. This leads to changes in their allelic frequencies. An attempt to capture the influences of these forces from genotypes is achieved by the generation of mathematical models. This is discussed below.

Referring to the genotypes in **Figure 1.4**, for simplicity, let us assume that in the first generation denoted by **t**, the genotypic frequency of parasites with the genotype **A** (represented by the **A** allele) is **p**, whereas the genotype frequency of **a** is **1-p**. If the two parasites mate to form a diploid, under random mating expectations, the genotypic frequencies of the diploids will be:

AA = p^2

Aa & **aA** = p (1-p) + p (1-p) or 2p (1-p)

aa = $(1-p)^2$

Following meiosis to restore haploidy, the expected frequencies of the alleles in the next generation **t + 1**, will be:

A = p^2 + 1/2 X 2p (1-p) = p^2 +p (1-p) = p^2 + p – p^2 = p

a = $(1-p)^2$ + ½ X 2p (1-p) = $(1-p)^2$ + p (1-p) = (1-p) [(1-p) + p] = 1-p

Thus if the allele frequencies in generation **t** are **p** and the frequencies in generation **t+1** are **p´**, **p** = **p´** whereby the genotypic frequencies remain the same from their initial values. Thus, the genotypic frequencies from one generation can predict the allele frequencies in the next generation and those allele frequencies can also be used to predict genotypic frequencies for the next generation, and so forth (adapted from Felsenstein, 2007).

1.5.2. Hardy-Weinberg equilibrium

The scenario described in the previous section reveals that the genotypic frequencies and the allele frequencies remain constant from one generation to another. This

forms the theoretical basis for the most commonly used model in population genetics known as the Hardy-Weinberg equilibrium (HWE). This model was generated using diploid organisms but since in this thesis the organism is haploid (with a brief diploid phase); the mathematical derivations have been adjusted accordingly.

This model was generated based on the following assumptions:

- The population in question comprises an infinite number of individuals
- No mutation
- The population is isolated with no migration
- There is no natural selection
- Sexual reproduction occurs and a zygote is formed by the random association of gametes.

This model forms the null hypothesis of population genetics studies such that departures from it i.e. deviations in genotypic or allelic frequencies across generations denotes that one of its underlying assumptions no longer hold.

1.5.3. Deviations of HWE expectations in natural populations

a) Genetic drift

In natural populations, the individuals that mate and contribute alleles to the next generation are of finite size. The result is that genotype and allele frequencies fluctuate from one generation to another due to sampling error. This stochastic event leading to changes in allele frequencies between generations is known as genetic drift (Ridley, 2004). For example, malaria parasites have been observed to cluster

around mosquito breeding sites (e.g. Midega *et al*., 2012), a situation implicating that they constitute a single subpopulation. In an area with numerous breeding sites and subsequently numerous subpopulations, genetic drift is expected to act on each of the isolated subpopulations (highly inbred) driving their alleles either to loss or fixation. The effects of genetic drift are higher in small population sizes than in larger such that fixation of these alleles occurs faster in the former than in the latter.

b) Mutations

Mutations are hereditary changes that occur in DNA sequences of organisms as a result of errors during replication and are the only source of genetic variation (Hamilton, 2009). Mutation is a slow acting process which barely alters the allele frequencies between generations.

In many parts of genomes it is speculated that, most mutations are non-lethal or neutral. The per site neutral mutation rate per generation (μ) in *P. falciparum* has been estimated to be approximately 4.91×10^{-9} substitutions per site per year or 6.88×10^{-9} substitutions per site per year assuming divergence from *P. reichenowi* (a parasite of chimpanzees) either of 5 or 7 million years ago (mya) (Ayala *et al*., 1998; Deirdre *et al*., 2003).

Point mutations play a significant role as they lead to either a single base substitution or to an insertion/deletion (indel) event in the genome. Single base substitutions constitute SNPs which will be used for analysis in this thesis (see **Chapter 3**). Indels lead to the addition or deletion of two or more nucleotides as a result of replication slippage, which sometimes disrupts the 3 codon amino acid reading frame. They

mostly affect repetitive loci such as microsatellites discussed further in this thesis in **Chapters 5 & 6**.

c) Migration

In natural populations malaria parasite populations are not isolated but they exchange genetic material with one another. The probability that two parasites will mate depends on their location within the population. It follows that parasites located within close distance of each other have a higher chance of mating than those farther away. This leads to population structure i.e. partitioning of a large population into sub-populations comprising closely mating individuals by virtue of their location in space.

Population structure leads to changes in genotypic and allelic frequencies by the partitioning of large populations into small genetically isolated subpopulations that are acted upon independently by evolutionary forces such as genetic drift or mutations. Deviations of HWE in this case occur due to non-random mating at a subpopulation level and genetic structuring at a population level. Further discussion of this is in **Chapters 5 & 6**.

(d) Natural selection

This is a force that affects genotypic and subsequently allelic frequencies. Natural selection acts on genotypes that have different viable phenotypes leading to their survival or loss across generations. The changes arising on the genotypic frequencies following natural selection lead to changes in allele frequencies.

In population genetic studies, loci under selection are not ideal in the interpretation of overall genetic structure of parasite subpopulations as the observed allele frequencies could be as a result of selection on these loci thus obscuring the neutral effects on allele frequencies due to genetic drift and gene flow patterns at a geographical level.

d) Non-random association of alleles (linkage disequilibrium)

LD is defined as the non-random association of alleles (Hamilton, 2009). To explain this, it is better to first define linkage equilibrium. When alleles located in loci of the same chromosome (unlinked) or in different chromosomes are genotyped and their allele frequencies calculated, due to genetic recombination and independent assortment of genes, it is expected that those alleles will occur at different allele frequencies. The occurrence of one allele is independent of the occurrence of the other, they are therefore randomly associated. These alleles are said to be in linkage equilibrium as expected under the classical Mendelian expectations. However, if these alleles keep occurring at the same allele frequencies than is expected by chance, due to linkage or due to their conferring a phenotypic advantage such that they are

inherited together, the occurrence of one allele is dependent on the occurrence of the other one in a non-randomly associated manner. They are therefore said to be in LD.

In *P. falciparum* the rate of recombination is high (as discussed in **Section 1.4.1**) such that LD is seldom observed. This is evidenced by a study of parasite isolates collected across Africa where linkage disequilibrium (LD) along the MSP1 gene was observed to rapidly decline with increasing nucleotide distance between the polymorphic loci analyzed. In fact, minimal LD was detected in regions located >1kb away from the MSP1 gene (Conway *et al*., 1999).

1.6. Project concept

The intended objective of this study is to trace the dispersal of malaria parasites in space based on their genetic relatedness. By using genetic markers, parasite genotypes representing the clones occurring within infections can be identified. The genetic relatedness (GR) of malaria parasites can then be calculated by pairwise comparison of the identified parasite genotypes using the equation:

$GR = 1-n'/n$

Where: n' = Number of SNPs not shared between the parasites

n = Total number of SNPs analyzed

The GR values can then be compared to the geographic distance of the parasite genotypes which is expected to be related to the distance the mosquito travelled to transmit the parasites (**see Figure 1.5**).

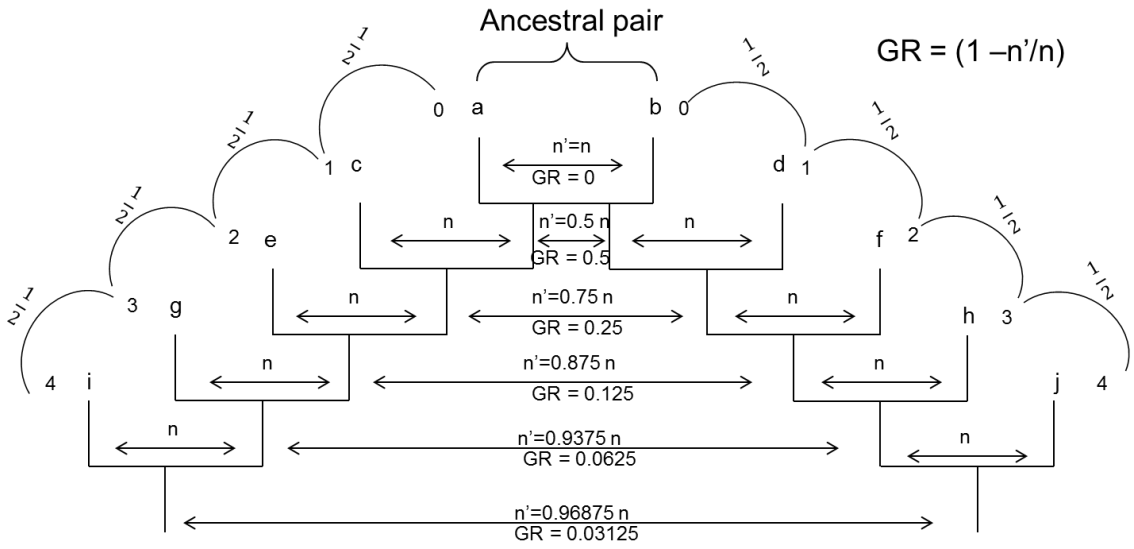


Figure 1.5 Schematic diagram showing the genetic relatedness of malaria parasites across different generations. The letters a-j represent different parasite clonal genotypes, n is the total number of SNPs analysed, n´ is the number of SNPs not shared between. The numbers 1-5 represent the generation number. Assuming random mating in the population, a & b are considered to be the ancestral pair (Generation 0) with no alleles shared between them thus GR = 0. When a mosquito picks the ancestral pair from one individual and they recombine after which they are transmitted to another individual, their immediate progeny (c & d) will have half of their alleles identical by descent to the ancestral pair as shown above. They will therefore have GR = 0.5 and the proportion of alleles not shared between them (n´) will be 0.5 of the total SNPs analysed (n). When this 1st generation progeny mate back into the population, their progeny i.e. e & f (2nd generation) will have a quarter of their alleles identical by descent to the ancestral pair (GR = 0.25), 0.75 of their total alleles not shared (n´), and they will share ½ of their alleles with the 1st generation (c & d). The 3rd generation (g & h) will have 1/8 of their alleles identical by descent to the ancestral pair, will share ¼ of their alleles with the first generation (c &d) and ½ of their alleles with the second generation (e & f), and so forth. Each generation represents the number of inoculations or transmission cycles from one individual to another by the mosquito. Thus from 0-1 generation that is one inoculation, from 0-2 generations, 2 inoculations and so forth. Parasites that are closely related to the ancestral pair will have high GR values. This means that there are fewer transmission cycles (inoculations) between them relative to the ancestral pair. However, parasites with low GR values define genetically distant clones relative to the ancestral pair representative of more transmission cycles between them (Figure developed by Carter, R).

In the representation depicted in **Figure 1.5**, the identified GRs of parasite clones can be used to determine the number of inoculations or transmission cycles between them relative to the ancestral parasites. If the geographic distance between the parasite clones is known, the distance the mosquito travels to transmit malaria parasites per transmission cycle can be predicted. For example, if two parasite clones having GR = 0.25 which is equivalent to 2 inoculations are collected 100m apart, that means for each inoculation, the mosquito has travelled 50m.

This information can then be used to trace malaria parasites in space as it will reveal the distance the mosquito travels between inoculations. The spatial patterns of malaria parasite distribution can then be resolved. Although attempts have been made to determine high malaria risk areas within small geographic settings, no study has so far used the genetic relationship of malaria parasites to reveal the distance of parasite dispersal between inoculations or per transmission cycle (as proposed in this section). If parasites are observed to be dispersed within a certain mean distance per transmission cycle, then the boundaries of malaria transmission can be established enabling the targeting of interventions to maximise their effectiveness. On the other hand, if the mean distance varies between transmission cycles with no established pattern of transmission, then different intervention strategies can be envisaged.

The concept described above forms the basis of the project design for this thesis. However, the work described in this thesis constitutes a feasibility study conducted prior to the actual testing of the above mentioned concept in natural parasite populations. The aims of the thesis are therefore outlined in the next section.

1.7. Aims of thesis

The study outlined in this thesis was devised to:

1. Identify suitable genetic markers (SNPs) across the *P. falciparum* genome for the identification of parasite clonal genotypes represented within a single blood isolate (**Chapter 3**).

2. Validate Pyrosequencing^{TM}, a technique which identifies and quantifies SNPs representing different parasite clones occurring in mixed clone infections (**Chapter 4**). This was done by analysis of laboratory prepared *P. falciparum* mixtures of clones represented at different proportions. In addition, the validated method and the identified SNPs were used in an analysis to determine the GR values in the recombinant progeny of a *P. falciparum* genetic cross of 3D7 and HB3.

3. Identify *P. falciparum* parasite genotypes in isolates collected from different geographic regions using the identified SNPs and microsatellites. This was done by the analysis of malaria positive isolates collected from individuals residing in Rusinga Island in Kenya, Kolle region in Mali and the South Western province of Cameroon. These isolates were collected in places separated by distances of <1km to >2000km. The obtained parasite genotypes were used in population genetic analyses such as: informativeness of the genetic markers as explained by levels of detected genetic diversity in the clonal genotypes, neutrality of the markers and the detection of population structure (**Chapter 5**).

4. Determine how widespread these parasites are in space by performing spatial autocorrelation tests (**Chapter 6**).

1.7.1. Hypothesis

In this study, the degree of genetic relatedness between identified parasite genotypes was determined. The hypothesis stating that closely-related malaria parasites are located within short geographic distances of each other will be tested. Based on the project concept described in **Section 1.6**, it is expected that closely-grouped parasites have fewer transmission cycles between them. The level of genetic relatedness is therefore expected to decrease with increase in geographic distance as this translates to more transmission cycles between the parasites by the mosquito from one individual to another.

The phenomenon that parasites at close geographic distance have a high chance of mating compared to those that are far apart, has been observed widely in other organisms and Wright, 1943 described it as isolation by distance. Its existence in natural populations governs the way individuals interact at a population level.

The fact that closely related parasites have a high chance of mating leads to non-random mating events, presenting as either inbreeding or partitioning of parasites into distinct subpopulations with limited gene flow between them (population structuring). In the case of malaria parasites, since mosquitoes cluster around breeding sites, transmission of malaria parasites by these mosquitoes would be confined around these sites. It is therefore expected that parasites closest to the breeding sites would represent a genetically distinct parasite population. However, since the transmission of malaria parasites is heterogeneous within small spatial settings i.e. between individuals residing in a single hut or individuals residing in

separate huts, further inbreeding, population structuring and limited gene flow is expected between the parasites.

On this basis it was hypothesised that closely-related malaria parasites collected from individuals residing within short distances of each other (~1km) exhibit high genetic relatedness and will therefore represent the shortest distance the mosquito travels per single inoculation compared to malaria parasites collected from individuals who are far apart (>1km and beyond).

2. MATERIALS AND METHODS

This study was devised to determine the genetic relationships of *P. falciparum* wild isolates and how this information can be used to define their interactions in space. Since these isolates exist in the form of mixed clones, prior to their analysis, it was pertinent to validate a technique that could enable the identification of the individual clones represented in each isolate. Pyrosequencing™ was the technique of choice for this study and this section describes the standardization of the technique.

2.1. Validation of Pyrosequencing™

2.1.1. Laboratory preparation of genetically distinct *P. falciparum* pure clones

The marker assays generated in **Chapter 3** were used to analyse *P. falciparum* field isolates that contain mixed clone infections for the identification of parasite genotypes represented within them and reveal the parasites' genetic relatedness. Before this was done, the technique and the selected markers were tested to ensure that the intended objective was achieved. This was done using laboratory prepared *P. falciparum* clones grown *in vitro* in RBC culture medium and harvested as pure clones.

The laboratory prepared samples (Parasitaemia 1-4%) containing different *P. falciparum* pure clones (3D7, CAMP, FCR3 or HB3) were used to determine which of the primer sets selected for analysis (see **Chapter 3**) gave the most accurate

allele (SNP) quantification scores (95-100%) using Pyrosequencing™. They were also used to confirm if the SNPs identified by the Pyrosequencing™ corresponded to the SNPs provided in the PLASMODB database in each locus The pure clones were prepared using uninfected whole blood at 5% hematocrit (packed cell volume of RBC present in whole blood), infected blood from culture at 5% hematocrit and 5% parasitaemia. These were grown for several days and then the parasite cultures were then serially diluted with uninfected whole blood at the same hematocrit in order to achieve 1-4% parasitaemias respectively.

The parasites were collected at such high parasitaemias to ensure good DNA yield for PCR analysis and subsequent analysis by Pyrosequencing™. These blood samples were then blotted on Whatman FTA® cards and left to dry. These cards are impregnated with a patented chemical formula that lyses cell membranes and denatures proteins on contact. Nucleic acids are physically entrapped in the cellulose fibres and can be recovered by washing away all cell debris and inhibitors of downstream analysis using standard wash protocols.

2.1.2. Laboratory preparation of *P. falciparum* samples containing genetically distinct clone mixtures

Having identified the allele specific primers that would amplify regions containing genetic markers for the identification of clones in mixed infections (**see Chapter 3**), the next step was to analyse samples containing a mixture of clones. Since *P. falciparum* mixed infections are common in nature, laboratory prepared clone mixtures were prepared to simulate what occurs in field isolates. This was done to test if the Pyrosequencing™ can identify and assign proportions to the clones

represented in the mixture enabling the identification of parasite genotypes represented in each mixture.

To accomplish this, *P. falciparum* laboratory prepared mixtures at different parasitaemia ranging from percentages of 0.1, 0.01, and 0.001 were analysed. The mixtures were prepared in proportions of 100:0, 98:2, 95:5, 90:10, 80:20, 50:50, 20:80, 10:90, 5:95, 2:98 and 0:100. The mixtures comprised *P. falciparum* cloned isolates of 3D7 and HB3, 3D7 and DD2, or 3D7 and 7G8 clones. The mixtures were prepared using uninfected whole blood at 5% hematocrit (packed cell volume of RBC present in whole blood), infected blood from culture at 5% hematocrit and 5% parasitaemia. The parasite cultures were then serially diluted with uninfected whole blood at the same hematocrit. A 20 µl sample was taken from each serial dilution and blotted on FTA® filter paper for subsequent analysis.

2.1.3. DNA extraction

(a) From samples derived from parasite cultures

To extract DNA, the malaria parasite culture suspension was spun down and the eluate discarded. It was then re-suspended to 10ml with PBS. DNA was extracted following the protocol provided in the Qiamp® DNA blood midi/maxi (Qiagen®). 500µl Protease K enzyme was pipetted into an empty 50ml centrifuge tube. The blood re-suspended in PBS was added into the 50ml tube together with 12ml buffer AL or the lysis buffer. This was thoroughly mixed and placed in a pre-heated 70°c water bath for 10 minutes to ensure maximum lysis of the cells for parasite DNA release. 10ml of ethanol was added thereafter and this was thoroughly mixed to

ensure maximum homogeneity of the solution for successful DNA precipitation. The mixture was then placed in a Qiamp® column provided in the kit. The column consists of a silica membrane that traps the precipitated DNA allowing protein and other contaminants that may contaminate the DNA to pass through. The column is then washed twice with buffer AW1 and AW2 provided in the kit. This ensures complete removal of contaminants from the bound DNA in the column. The DNA was then eluted with buffer AE (elution buffer). The DNA was finally collected and stored at -20°c for analysis by Pyrosequencing™.

(b) From samples blotted on Whatman FTA® cards

To prepare a sample for analysis, a disc was taken from the dried blood spot using either a 1.2 mm or 2.00 mm Harris micro punch and placed in a 0.2 ml PCR amplification tube. It was taken through three washes using 150µl FTA® purification reagent with vortexing for 10 seconds in between the washes to ensure complete removal of blood which may contain nucleases or proteases that may compromise the subsequent analysis of the DNA blot by PCR. It was finally washed off with Milli-Q pure water (Millipore) then air dried. It was then transferred to a 0.2 ml PCR tube ready for analysis under standard PCR conditions.

2.1.4. PCR analysis for Pyrosequencing™ assay

Each locus was amplified in a 50 µl volume using the following reagents and reaction conditions: 1x immoBuffer, 1.5 mM $MgCl_2$, 150 µM dNTPs, 1 unit Immolase™ DNA polymerase (Bioline) and 0.4 µM for the forward and reverse

primer respectively. The previously prepared disc in a 0.2 ml PCR amplification tube with immobilised DNA was used for the PCR reaction.

The reactions were performed in a DNA thermocycler (Biometra, T3000.), and programmed as follows: First denaturation step at 95°C for 7 min, first annealing at 42°C for 1 min, first extension at 72°C for 2 min, second denaturation at 95°C for 3 min, second annealing at 44°C for 1 min, followed by a total of 38 amplification cycles comprising: denaturation at 95°C for 1 min, annealing at 49°C for 1 min and extension at 72°C for 2 min. The reaction was ended by a final extension time of 5 min at 72°C. PCR products were visualised on 1 % agarose gels and samples that produced single amplicons of the predicted length were used for subsequent Pyrosequencing™ analysis.

Negative and positive controls were included to ensure no contamination of the PCR as well as to validate the results. The PCR protocol is shown in **Appendix A**. Results indicated that some of the 1.2 mm discs containing DNA failed to yield a PCR product hence the 2.0 mm discs were ideal.

2.1.5. Pyrosequencing™ assays

These were carried out on a PSQ™ HS-96A instrument. The single stranded PCR products served as the template for the Pyrosequencing™ analysis. One of the primers was biotin tagged; therefore the PCR products are biotinylated. The biotin tag renders the DNA ability to bind to streptavidin beads and remain immobilised on them. The unbound DNA strands are washed off. The sequencing primer is allowed

to hybridise to the template DNA and the sequencing reaction proceeds in the PSQ™ HS-96A instrument.

To start the Pyrosequencing™ analysis, two plates were set up (Plate 1 and Plate 2). Each well in Plate 1, a clear and round-bottomed 96 well plate, was loaded with 40 µl high purity Milli-Q water (Millipore), 38 µl of binding buffer (Biotage), 2 µl of Streptavidin Sepharose high performance beads (Amersham Biosciences, Sweden) and 2-4 µl of DNA (dependent on band intensity of gel photo). Some wells were set up as negative controls devoid of DNA to check for contamination. Plate 2 (PSQ™ HS-96-well, Biotage), an opaque flat bottomed 96 well plate was loaded with 11.5 µl annealing buffer and 0.5 µl sequencing primer (10 µM final concentration). Plate 1 was placed on a shaker to allow binding of the biotinylated single stranded PCR product to the streptavidin beads.

Using a Pyrosequencing™ vacuum prep tool (Biotage) containing probes to fit into each well of Plate 1, the DNA immobilised onto streptavidin beads was sucked and suspended onto the probes. The probes were then immersed in 70% ethanol wash buffer for 5s and then onto 0.2M Sodium hydroxide for 5s to remove the non-biotinylated unbound DNA strands. The final wash was done in a wash buffer (10mM Tris acetate, PH 7.6) for another 5s. The vacuum was then switched off to release the DNA into Plate 2. The plate was incubated at 80 °C for 2 min to allow the sequencing primer to hybridise to the DNA and then allowed to cool before transfer to the PSQ™ HS-96A instrument. Analysis was then performed in the PSQ™ HS-96A instrument using enzyme, substrate and nucleotides from Biotage.

2.2. Preparation of P. falciparum genetic cross of 3D7 and HB3

Sexual recombination in *P. falciparum* was investigated in the laboratory following the generation of a genetic cross containing the genetically distinct clones 3D7 and HB3 (Walliker *et al*., 1987). The progeny of this cross was tested using various genetic markers characteristic of each of the parental clones. These tests confirmed that sexual recombination had indeed taken place as the various recombinants recovered contained both parental genomes represented in them (Walliker *et al*., 1987). Material from this original 3D7 and HB3 cross progeny was grown in culture containing complete RPMI at 5% hematocrit.

In natural *P. falciparum* infections, an individual is typically infected by 1-3 genetically distinct clones. Using the assumption that the progeny distribution in the RBCs follows a Poisson expectation (**Appendix B**), the calculations outlined below were carried out to reduce the number of recombinants in the culture to a level that mimics the number of clones infecting individuals in natural *P. falciparum* malaria infections i.e. 1-3 recombinants per diluent. This process of reducing the number of recombinants by limiting dilution is referred to here as semi-cloning.

2.2.1. Thawing of cryopreserved cross progeny material

The culturing process was begun by thawing the cross progeny material in liquid Nitrogen storage. Solutions for use in this process prepared under sterile conditions comprise: A. 12% Sodium chloride, B. 1.6% Sodium Chloride & C. 0.9% Sodium chloride + 0.2% dextrose). The cryopreserved vial was removed from the storage and placed in a 37°C water bath until fully thawed. The blood sample was then

transferred to a sterile 50ml falcon tube after its volume was estimated its volume. Using 0.2ml of solution A per 1m of the thawed blood sample the solution A was added slowly drop by drop with constant mixing but gently. The tube was allowed to stand for 3 minutes. This was done to ensure that the frozen parasites were recovered with minimal loss due to lysis of the RBCs. 10ml Solution B for every 1ml of the original thawed sample was then added drop by drop with gentle continual mixing. Thereafter, Solution C was added drop by drop using 10ml for every 1ml of original blood sample. Centrifugation was then performed at 2000rpm for 5 minutes. The supernatant was removed.

The cells recovered were then re-suspended slowly in complete medium at 5% haematocrit then cultured. The complete medium was prepared using 62.25ml RPMI 1640 (containing 20mM Sodium Bicarbonate and 25mM HEPES buffer. Addition of 5m glutamine (2mM concentration), 5ml of 20% glucose solution, 25μg/ml gentamicin, 1ml of Sodium Hydroxide at 7.2-7.4 PH, and 40ml human serum followed thereafter. Since the complete medium was stored at 4°C, if added directly to the parasite culture at this temperature, parasites would die of cold shock. It was therefore warmed in a water bath at 37°C before use. 10ml of the suspension was transferred to a $25cm^2$ flask. Gas mixture containing 1% oxygen, 3% Carbon Dioxide and 96% Nitrogen was added for 60 seconds and the flask cap screwed tightly to avoid gas loss. The flask was then placed in an incubator maintained at 37°C.

2.2.2. Parasite growth in culture

Daily monitoring of the culture is recommended by preparation of blood smears to ensure that the parasites are thriving. In addition, gassing is also performed to ensure optimal parasite growth conditions. Parasite smears were prepared by taking a small drop from the flask and placing in a 500µl eppendorf tube then spun down by quick spin centrifugation. The supernatant was removed and the pellet resuspended to 50% haematocrit using complete medium. This was then smeared onto a clean glass slide, fixed with methanol to ensure the cells do not slough off when stained with 10% Giemsa. The Giemsa was prepared using Phosphate buffered saline (PBS) at PH 7.2. This was then viewed under a microscope with immersion oil at objective X100. Parasitaemia was estimated by counts of approximately 500 cells. A drop of blood was added on the 2nd day of culture coinciding with the bursting and reinvasion of ring stages to fresh RBCs.

The blood for culture maintenance was obtained from the Scottish National Blood Transfusion Service. The blood is drawn into packs containing the anticoagulant Acid Citrate Dextrose (ACD) or Citrate Phosphate Dextrose (CPD) but not heparin as it hampers parasite growth. The blood was spun down by centrifuging at 2000rpm for 5 minutes. The serum was aspirated off followed by 3 washes with centrifuging and re-suspension using incomplete RPMI 1640 incomplete medium (lacking the human serum). 100% of the incomplete medium or more of the volume of cells was used in the washes. The RBCs were then re-suspended at 50% haematocrit in complete medium.

To ensure efficient growth of the parasites, blood and culture need to be added. To know the amount of blood needed, the parasitaemia was first estimated. Since most parasites give approximately a 5 fold reinvasion each cycle, the amount of blood added depended on the needed parasitaemia. For example, starting with a 200µl packed cell volume (PCV) at 10% parasitaemia, a 1:5 dilution was achieved by adding 800µl PCV fresh RBC which reduced the parasitaemia to approximately 2%. Thus the next reinvasion cycle should take the parasitaemia back to about 10%.

The equation used to estimate the amount of complete medium to add when changing the medium during the culturing process was:

Medium (ml) = 5 X % parasitaemia X PCV (ml)

The culture was first transferred to a 15ml falcon tube for a 4 minute centrifugation at 2000rpm. The supernatant was then removed comprising of the old culture. Fresh medium was then added with gentle mixing. The culture with the fresh medium was then transferred to 25cm^2 sterile flasks gassed and incubated at 37°C.

2.2.3. Parasite synchronization

The parasite culture was synchronized to ensure that the parasites were at the same growth stage (rings) for the limiting dilution assay. Synchronization of the cultures was achieved by growing the parasites to ~1% with a high percentage of ring forms and re-suspending to original volume using sorbitol (made up of 5% sorbitol in PBS). This was allowed to stand at room temperature for 5 minutes and then centrifuged. The sorbitol solution was removed and the culture was re-suspended in complete RPMI medium and washed twice. After washing, the culture was re-suspended in

complete medium and a smear taken after 24 hours to ensure that synchronization was successful. This was grown for a week and at a parasitaemia of ~7.7%, the number of parasitized red blood cells (RBC) was determined by performing an RBC count using a haemocytometer to estimate the total RBC count in the culture which was found to be $4.9x10^8$ RBC per millilitre (ml).

2.2.4. Semi-cloning the 3D7 and HB3 cross progeny

No. of parasitized RBC (pRBC) per ml: $7.7/100 \times 4.9x10^8 = 3.8x10^7$

That means there is $3.8x10^5$ pRBC per 0.01ml (10µl)

10µl of culture was made up to 3.8ml of complete RPMI medium containing fresh washed blood at 5% hematocrit. More RPMI complete medium containing freshly washed blood at 5% hematocrit was prepared to make up the suspensions shown below.

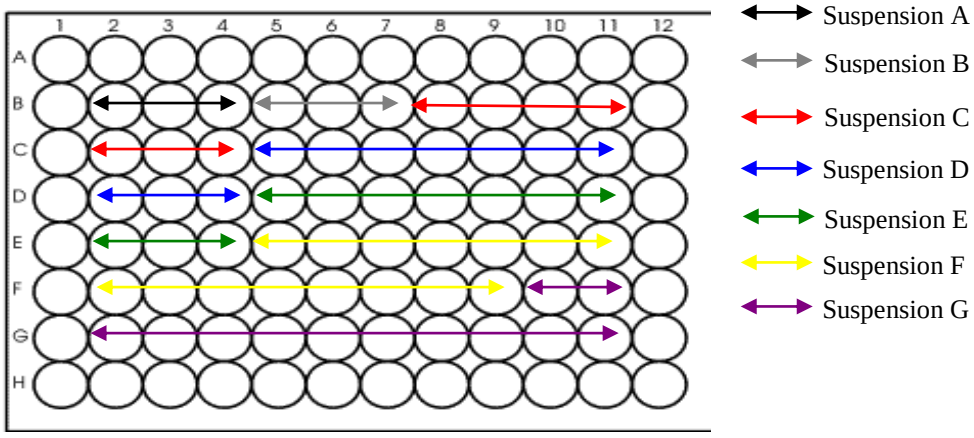
There is now $1x10^5$ pRBC in 1ml

1. $1x10^3$ pRBC in 10µl. This was called Suspension **A**
2. 100µl of Suspension A was made up to 1ml; 100µl of this suspension contained 10^2 PRBC……Suspension **B**
3. 200µl of Suspension B was made up to 2ml thus 100µl of this suspension contained 10 pRBCs.....Suspension **C**
4. 100µl of Suspension C was made up to 2ml; 100µl of this suspension had 5pRBC……..Suspension **D**

5. 40µl of Suspension D was made up to 2ml therefore there were ~2.5pRBC in 100µl……..Suspension **E**
6. 140µl of Suspension E was made up to 2ml and this had ~0.7pRBC in 100µl……. Suspension **F**
7. 40µl of Suspension F was made up to 2ml giving ~0.375PRBC in 100µl…….. Suspension **G**

The suspensions above were taken and set up in a plate (**Figure 2.1**)

(a)



(b)

Suspension	Calculated number of pRBCs	No. of wells in plate
A	10^3	3
B	10^2	3
C	10	7
D	5	10
E	2.5	10
F	0.7	15
G	0.375	12

Figure 2.1 Schematic diagram showing the plate set-up for the 3D7 and HB3 cross progeny semi-cloning by limiting dilution. **(a)** Shows how each of the diluents were set up in the plate to culture. Each of the colours indicates the suspension contained in each well. The outer wells were loaded with complete RPMI to ensure that the cultures do not get dehydrated. Note that the number of wells increases as the number of pRBC per suspension decreases to increase the chances of detecting parasites per well. **(b)** Table showing a summary of the suspension, the estimated number of pRBC in each suspension and the number of wells containing each suspension.

The cultures in the plate were grown for 2 weeks and smears taken every day to monitor parasite growth. All wells in A, B, C were found to be positive, 8 out of the 10 wells in D containing estimated 5 pRBCs per well were positive, while all wells in E, F and G remained negative. From the Poisson distribution, ~80% positive wells are expected to occur in wells estimated to contain ~1.6 pRBCs and 99-100% positive wells in those estimated at ~5 pRBCS (**Appendix B**)

The fact that 8 out of 10 (80%) of wells were positive in row D, (estimated to have 5 pRBC per well) indicates an average of ~1.6 clones per well. This suggests, according to the Poisson distribution curve (**see Appendix B**), that there was either an over-estimation in the calculations or a loss of parasites during the culturing process. All the cultures in the D wells were taken and transferred into bigger flasks to grow them and increase parasitaemia for higher DNA yield. The parasites were harvested at parasitaemias of between 6-15% and DNA extracted from them.

2.3. Analysis of field isolates

Isolates were collected from Cameroon lying approximately 2,300km from Mali and approximately 2,800km from Kenya, and, Mali lying approximately 5,100 km from Kenya (**Figure 2.2**).

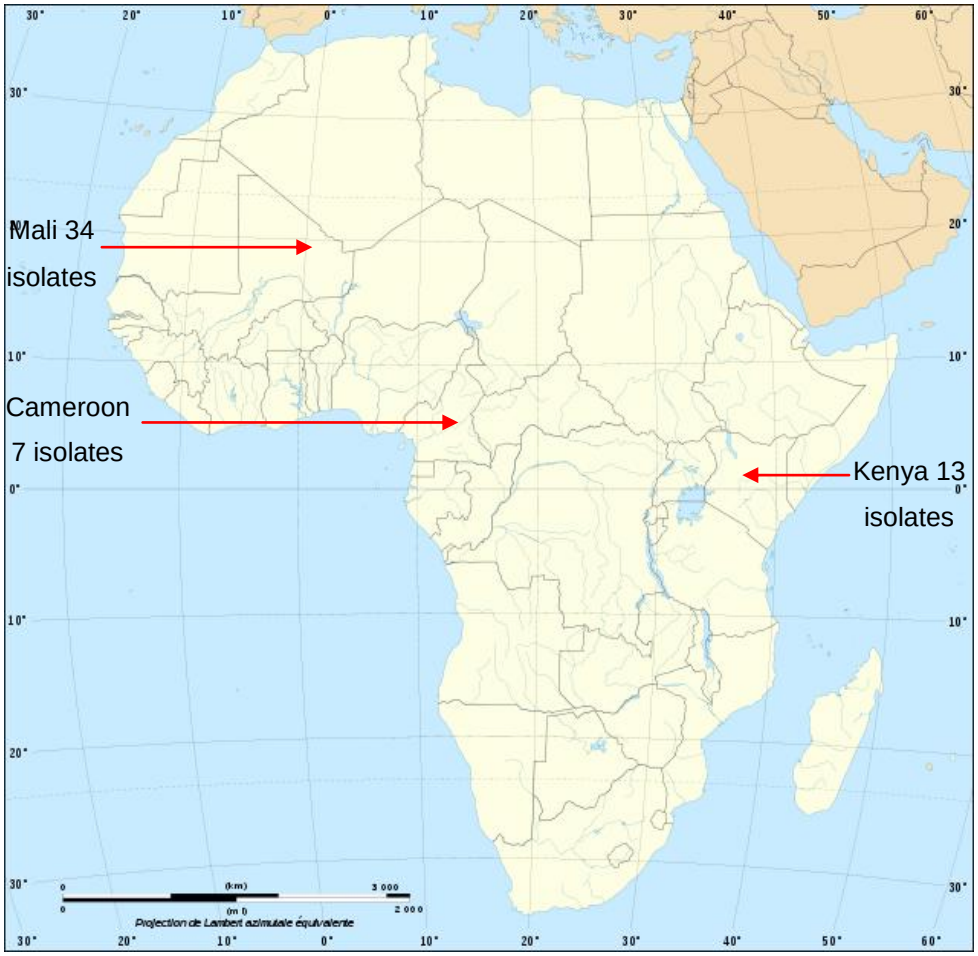


Figure 2.2 Map of Africa drawn to scale and showing the countries from which the isolates were collected.

2.3.1. Sample collection in Kenya

2.3.1.1. Study site

Blood samples were collected from Rusinga Island which has an area of 42 km^2 and is located in Suba District, Nyanza Province, Western Kenya, within Lake Victoria, on the southern shores of the Winam gulf of the lake. Due to its close proximity to the mainland, a 200 m long causeway was constructed in 1983 to link the island with

Mbita Township, the major trading and administrative centre of Suba district (**Figure 2.1**).

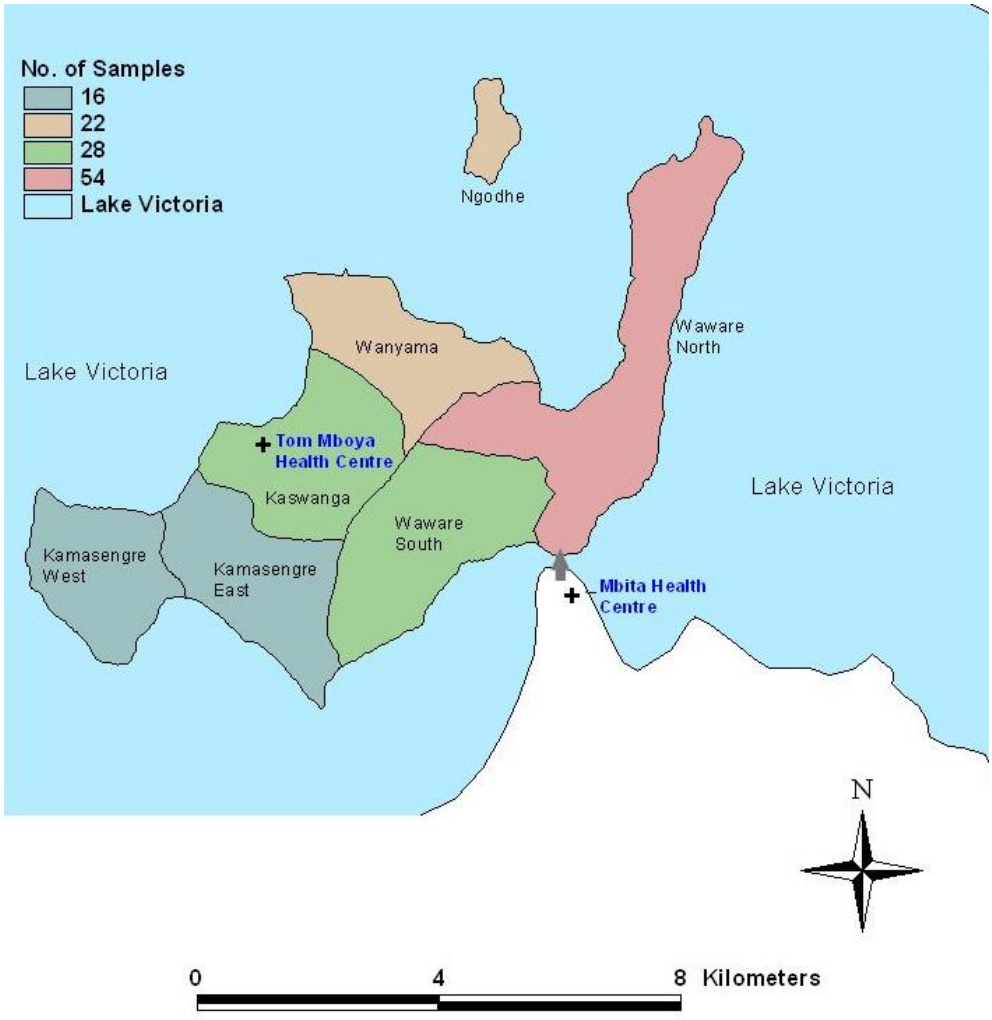


Figure 2.3 Map of Rusinga Island showing part of Mbita that is in the mainland (in white) connected to Rusinga Island by the 200m causeway (grey arrow). The island is demarcated into large locations as shown in this map. Malaria positive samples collected from each location are shown. During the time of the sample collection, there was no clear demarcation of the Island thus Kamasengre East and West were merged (the area was known as Kamasengre). Kaswanga and Waware South were also merged (the area was known as Kaswanga) (ESRI ArcView version 3.2 - http://www.esri.com/)

Two rainy seasons are typical for the area, the 'long rains' between March and June and the 'short rains' between October and November. These seasonal patterns are

unreliable with some years characterised by prolonged dry periods and others by prolonged wet seasons. Malaria transmission fluctuates seasonally but is sustained all year round by three primary malaria vectors *Anopheles gambiae, An. arabiensis* and *An. funestus*; (Chen *et al.,* 2004 and Minakawa *et al*., 1999) Mosquito breeding sites in this area include: man-made water collection points such as open cemented tanks, pools made from animal hoof prints, ponds, marshes (Fillinger, 2004). In addition, the water levels of the lake have been receding leaving a patch of land that has been identified as a potential breeding site (Minakawa *et al*., 2008).

A census implemented by the end of 2006 during the establishment of a Demographic Surveillance System (DSS) indicated that Rusinga island had 24,078 inhabitants which formed 5,425 households and 21 % of the population were children below the age of five (Satoshi Kaneko, Nagasaki University, unpublished data).

2.3.1.2. Study Population

120 malaria positive samples were collected (**Table 2.1**). The sample size was chosen on the basis of adequacy in estimating allele frequencies. Since most of the population genetic studies are pegged on allele frequency numbers, a sample size giving good representation of allele frequencies for the whole *P. falciparum* population is favoured. Inclusion criteria were:

a) Malaria positive individuals of all ages
b) Presenting with fever and other symptoms consistent with malaria e.g. vomiting, lack of appetite, convulsions.
c) Blood smears showing *P. falciparum*
d) Inhabitants of Rusinga Island. At the time of sample collection the actual geographic distances of separation for the sample collection were not taken into consideration. Thus for this study site the geographic coordinates specifying the actual residence of the inhabitants is missing.

2.3.1.3. Ethical Clearance

The institutional ethical clearance was granted by the joint University of Nairobi - Kenyatta National Hospital ethical review committee (protocol approval number P156/9/2005) (**Appendix C**). In addition, permission was obtained from the district authorities. Blood samples were collected only after individual or parental consent.

2.3.1.4. Study design

Rusinga Island has been reported to have a high prevalence (50%) of malaria (GoK reports, 2001). However, this has reduced conceivably due to intervention strategies such as usage of nets as well as availability of drugs in the health centre. Further, there have been changes in weather patterns with extended wet periods that may lead to washing away of mosquito larvae hence low malaria transmission by the vectors.

In addition, over the counter drugs are readily available so individuals are recorded to have taken them prior to visiting a health centre making it difficult to detect malaria parasites by microscopy.

Initially, blood samples were collected from individuals reporting for treatment and falling within the inclusion criteria irrespective of whether they were malaria positive or negative, by a single prick using a lancet about 2 mm behind the nail bed. The flowing blood was then blotted onto the Whatman FTA® cards and allowed to dry completely before storage in a desiccant (silica) ready for shipping to the University of Edinburgh. Some blood was also placed on a glass slide and thin and thick smears prepared on the same slide and left to dry. The thin smear was fixed with 99 % methanol and left to dry completely. The smears were then stained using Giemsa's stain and stored for later analysis by light microscopy.

However, since few malaria positive cases were reported and due to time constraints, the sample collection process was reviewed and extended to individuals over 5 years and malaria positive cases confirmed by microscopy. This involved making a prick first to determine whether the patient was malaria positive or negative by light microscopy. Once identified as malaria positive, then a second prick was made and blood was blotted onto the Whatman FTA® cards.

Because of insufficient numbers of samples collected from the villagers attending the clinic, there was also need to visit the nearby schools and perform mass free malaria screening and medical campaigns in order to increase the probability of detecting malaria positive cases.

2.3.1.5. Data collection

Samples were collected between May and October, 2007 the 'long rains' period.

Blood samples were collected from:

1. Mbita Sub-district Hospital located in Mbita Township which is the major trading and administrative centre of Suba district. Malaria positive samples were obtained from individuals attending the medical facility for treatment.
2. Tom Mboya Memorial Health Centre (TMMHC) situated on the North-Eastern part of Rusinga Island. Malaria positive samples were collected from individuals attending the medical facility for treatment.
3. Schools located in Rusinga Island namely:
    a) Kibisom Nursery School located approximately 100m from the TMMHC in Kaswanga Location, Rusinga Island. 20 children aged 3-6 years old were chaperoned to the health centre and screened for malaria. No malaria positive cases were reported from this school.
    b) Agiro Primary School situated approximately 1 km north from the TMMHC in Kaswanga location, Rusinga Island. Screening was conducted within the school premises. 209 children aged 1-15 years were screened; of these 13 were positive for malaria. These were taken to the TMMHC for medication.
    c) Kaswanga Primary School Located approximately 1 km south from the TMMHC. Screening was conducted within the school premises. 116 children aged 1-15 years were screened; of these 4 were positive for malaria and were treated at TMMC.

d) Island of Hope Humanist Centre situated approximately 5 km from the TMMHC in Kamasengre location, Rusinga Island. Screening was conducted within the school premises. 145 children aged 1-15 years were screened; of these 6 were positive for malaria and were transported to TMMC for treatment.

2.3.1.6. Determination of Parasitaemia

Thin and thick blood smears were prepared on one glass slide and left to dry. The thin films were fixed with methanol and both stained with Giemsa's stain. These were later analysed by light microscopy. Parasite density was assessed by counting the number of parasites per 200 leukocytes in an oil immersed thick film (using a microscope with X100 objective and X10 eyepiece).

A sample was considered negative after 20 high powered fields had been read without parasites. Although the protocol demands reading of 200 high powered fields, only a maximum of 20 fields could be read due to time constraints. Parasite counts were converted to parasites per 1 μl blood assuming an average of 8000 leukocytes per μl of blood for thick films. For the thin films, the number of RBC present in the average microscopic field was determined as it varies depending on anaemia occurrence in an acute infection. The total RBC count value of $5.0X10^6$ RBC/μl (males) and $4.5X10^6$ RBC/μl (females) is generally assumed (Shute, 1988). The number of asexual parasites per 25 microscopic fields was counted and the total parasite count per μl of blood calculated. The parasitaemia ranged from forty to three hundred and fifty six thousand parasites per microliter of blood (Thick smear).

Table 2.1 Summary of *P. falciparum* infected blood samples collected from Rusinga Island, Kenya

Location	Male	Female	**Total**
Kamasengre East and West	10	6	**16**
Kaswanga & Waware South	16	12	**19**
Wanyama & Ngodhe	14	8	**22**
Waware North	25	29	**54**
Total	**65**	**55**	**120**

Of the 120 samples collected from this region, only 15 samples were analysed in this study. These samples were selected on the basis of high parasitaemia levels. The rest of the samples were not included in this feasibility study which serves as a proof of method and whose preliminary data will inform a larger planned future study with the remaining isolates. In addition this study had budgetary constraints plus the missing geographic co-ordinates hence the analysis of the few samples. These samples were collected from patients attending health clinics and information on their place of residence was not collected. Nevertheless the geographic co-ordinates of the locations (given in **Table 2.1**) were used in the analysis in this study.

2.3.2. Sample collection in Mali

2.3.2.1. Study site

P. falciparum malaria positive samples were collected from Mali in a region 60 kilometres south of Bamako (Mali's capital) known as Kolle. In this region, *P. falciparum* malaria has a prevalence of 40-50% during the low peak season (October

–May) and 70-85% in the high peak season (June-September). The major malaria vectors in Mali are: *Anopheles funestus, An. brohieri, An. gambiae, An. flavicosta, An. coustani, An. pharoensis, An. arabiensis, An. nili, An. hancocki,* and *An. paludis* (WHO report, 2006). Kolle has about 2,500 inhabitants with 69 families living within a radius of 600 metres from the nearest female A*nopheles* mosquito breeding site (River Kolle) and 5 other families within a radius of 1000 metres (**Figure 2.2 a & b**).

(a)

(b)

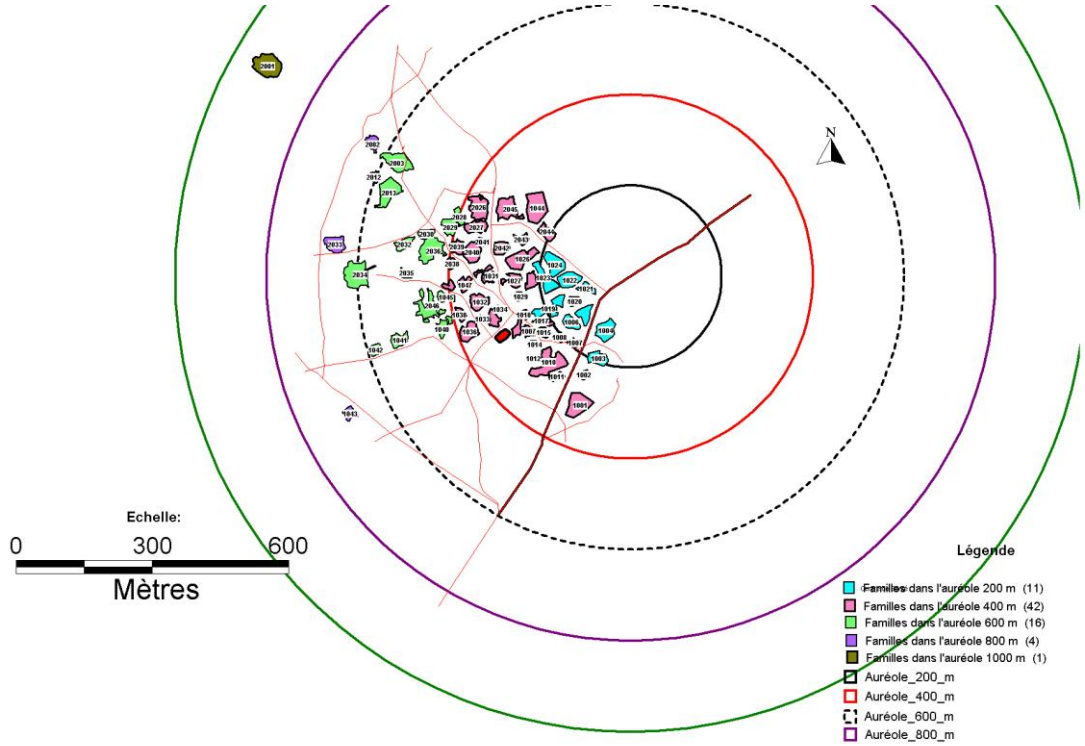


Figure 2.4 Maps showing (a) Mali with study site (Kolle) indicated in red and (b) Kolle region divided into 5 aureoles increasing in radii by 200 metres and the location of the households with household numbers outlined in green, blue, pink, purple and dark green . Demarcating by aureoles was established by the group in charge of this project for easy sampling and identification of households represented in this area. The River Kolle, which is the main mosquito breeding site in this region transverses across the study site (maroon line).

2.3.2.2. Study population

P. falciparum malaria positive blood samples were collected from 51 individuals of all ages (**Table 2.2**). This sample collection was done between September – December2010. The lack of geographic coordinates in the Rusinga Island sample collection necessitated additional sample collection from this study site. Thus the

actual residences of individuals from which the samples were collected are known. Further to this, the study site is <1km enabling determination of *P. falciparum* spatial patterns in a region with short geographic separation distances. In addition, individuals enrolled in the study had to fill a questionnaire containing the following variables:

1. Age
2. Sex
3. Ethnicity
4. Household identification number (ID)
5. Head of household name
6. Village and any additional information that can assist in tracing the house if the household ID gives problems
7. List of symptoms:
    a. fever
    b. anaemia
    c. Other signs of severe disease
8. Last reported malaria attack
    a. Date or month of malaria attack
    b. Whether it was diagnosed at the clinic or elsewhere
    c. What treatment was taken

2.3.2.3. Ethical clearance

This was granted by the ethical review Board of the Faculty of Medicine, Pharmacy and Dentistry, Bamako, Mali (ethical review number 0940/FMPOS). Blood samples were collected following individual or parental consent.

2.3.2.4. Study design

Blood samples were collected from individuals reporting to the clinic (located in aureole 3, **Figure 2.2b**). Following a needle prick in the finger, blood was collected on a slide for a thick smear for malaria parasite detection. The slide containing the

thick smear was stained with Giemsa's stain and parasitaemia determined by microscopy as previously described. Thus, 51 malaria positive blood samples with parasitaemia between 1075 - 74,550 parasites/μl were collected and blotted onto Whatman FTA® cards. These were left to dry then stored in pouches containing desiccant ready for shipping to the University of Edinburgh.

Table 2.2 Summary of *P. falciparum* infected blood samples collected from Kolle, Mali

Aureole	**Male**	**Female**
1	3	8
2	13	10
3	5	10
5	1	1
Total	**22**	**29**

2.3.3. Sample collection in Cameroon

2.3.3.1. Study site

Cameroon is located in the Central West Africa region. The study site, Buea is located in the South West province of Cameroon on the lower slopes of Mount Cameroon (**Figure 2.3**).

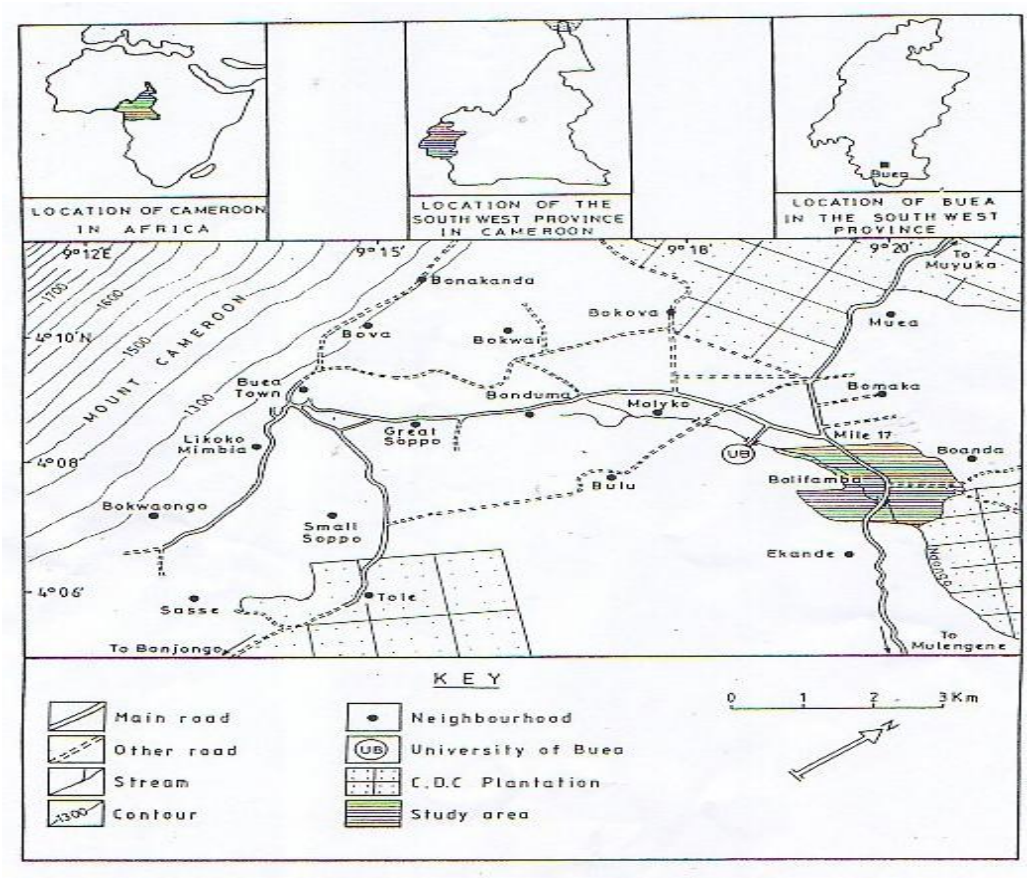


Figure 2.5 Map of Cameroon's South Western province showing the locations where the blood isolates were collected i.e. Buea, Molyko and Muea. (Map provided by Anong D, collaborator in Cameroon).

P. falciparum malaria transmission is perennial in this region and is mainly transmitted by *Anopheles gambiae sensu stricto*. The high malaria peak transmission season is during the rainy season between March and October and the low peak season is during the dry season between November and February. The main mosquito breeding sites include: fish ponds, man-made cemented tanks, hoof prints, roadside ditches (Wanji *et al*., 2009).

2.3.3.2. Ethical clearance

The study was approved by both the University of Buea's ethical review committee and the Cameroonian South West Regional Delegation of health (ethical review number R11/MPH/SWP/PDPH/FP-Research/5077/84). Blood samples were collected only after individual or parental consent.

2.3.3.3. Study Design

13 *P. falciparum* malaria positive blood samples were collected from individuals residing in Buea region who reported to the local health post (**Table 2.3**). The inclusion criteria was individuals of all ages with parasitaemias >20,000 parasites/μl and who resided in close proximity to each other (within a radius of 5km). The samples were collected within the month of September. The reason for this sample collection was to analyse them by Illumina Solexa sequencing. However, preliminary analyses of the 3D7/HB3 cross progeny by Solexa sequencing yielded numerous non-specific sequence reads following alignment with the reference sequence, 3D7. This may be caused by either human DNA contamination or the misalignment of non-unique sequences. This is a major setback especially when dealing with multiclonal infections. The strategy was discarded but the samples were analysed using Pyrosequencing^{TM}. The geographic coordinates for the residences of individuals from which these samples were collected are also missing for this study

site. However, geographic coordinates for Molyko, Buea and Muea were used in the final analysis outlined in this thesis.

2-4 ml of blood was collected into a labelled tube containing the anticoagulant Acid Citrate Dextrose (ACD). The tube was inverted several times to allow mixing with the anticoagulant. The blood was then spun down by a bench top centrifuge at 2500 revolutions per minute (rpm) and the plasma and buffy coat layer were removed.

The red blood cells (RBCs) were then washed twice with 10ml Phosphate buffered saline (PBS) buffer with centrifugation and supernatant removal per wash. The RBCs were then resuspended in 10ml PBS solution and transferred to a sterile 50ml tube. The RBCs were spun down and the supernatant removed. The packed cell volume (PCV) was estimated. Glycerolyte was then added at a ratio: 3 volumes of PCV to 5 volumes of glycerolyte. The glycerolyte was added to the RBC pellet slowly and drop wise with continual gentle swirling of tube to ensure even mixing with minimal RBC lysis. The RBC/glycerolyte mixture was then transferred to Nunc^{TM} cryotube freezing vials and stored at -70°C then to liquid nitrogen the next day. The samples were then stored in liquid nitrogen ready for shipping to the University of Edinburgh.

Table 2.3 Summary of *P. falciparum* infected blood samples collected from Cameroon

Region	Male	Female	Total
Buea Town	0	2	2
Molyko	2	3	5
Great Soppo	1	0	1
Mutengene	0	1	1
Bomaka	1	0	1
Bolifamba	0	1	1
Bova	0	1	1
Sosse	1	0	1
Total	**5**	**8**	**13**

2.3.3.4. Thawing of frozen blood samples and culturing

The frozen *P. falciparum* malaria positive samples were stored at -70°C at the University of Edinburgh. 8 of the blood samples were thawed by removing the Nunc™ cryotube freezing vials from the -70°C freezer and transferring to a 37°C water bath. Each thawed blood sample was transferred to a 50ml tube. Sterile 12% NaCl was added dropwise and slowly with gentle shaking to minimise RBC lysis at a ratio of 200µl NaCl to 1ml thawed blood suspension.

This suspension was allowed to stand for 5 minutes then 10ml of 1.8% NaCl solution was added slowly and drop by drop. Finally, 10ml of 0.9% NaCl plus 0.2% glucose was added. The suspension was then spun down by centrifugation for 4 minutes at

2000 rpm. The supernatant was aspirated and the RBCs resuspended in 20ml incomplete RPMI. This was spun again as previously described and the supernatant removed.

The PCV was determined and the RBC pellet resuspend in complete RPMI at 2% hematocrit. A small aliquot was taken and was used to prepare a thin smear for parasitaemia determination by microscopy. The suspension was transferred to a culture flask then gassed with a mixture of 94% Nitrogen, 5% Carbon dioxide and 1% Oxygen for 30-60 seconds. The culture was incubated at 37°C for 24 hours until the ring stage parasites matured to schizonts or reinvaded new RBCs for enrichment of DNA required for subsequent analysis. The parasitaemia prior to culture was determined and recorded.

2.3.3.5. Extraction of DNA

After 24 hours of culture, an aliquot of the blood sample was taken for parasitaemia determination by microscopy. Once the ring stage parasites were observed to have matured to trophozoites or reinvaded new RBCs, the culture was taken and DNA extracted using the Qiagen® kit as previously described.

2.4. Computer packages and software

Pyrosequencing™ assays were designed using the PSQ™ assay design software (**Chapter 3**). The data presented in **Chapters 4, 5 & 6** was processed and stored in

Microsoft office Excel 2003 for easy management. Figures and tables were generated by the Microsoft office Excel 2007. Analysis of SNPs and microsatellites was carried out using the software FSTAT vs 2.9.3 (Goudet, 2001) (**Chapter 5**). The spatial autocorrelation analyses of the combined data set in **Chapter 6** were perfomed using the Microsoft Excel, add-in software GenAlEx 6.5 (Peakall and Smouse 2006; 2012).

The map of Rusinga Island, Kenya was generated using the mapping software distributed by the Economic and Social Research Institute (ESRI ArcView version 3.2 - http://www.esri.com/).

3. GENERATION OF MARKERS AND PYROSEQUENCING™ TECHNIQUE VALIDATION FOR INVESTIGATION OF GENETIC RELATEDNESS OF *P. FALCIPARUM* FIELD ISOLATES

3.1. Introduction

The genetic composition of parasites in malaria infections can be investigated using genetic markers such as allozymes, single nucleotide polymorphisms (SNPs) and microsatellites (see **Chapter 1**). These markers help identify parasites that are genetically related in a mixed clone infection by identifying parasites that carry the same gene allele combinations (genotypes). Various molecular methods have been developed to identify these markers and have been used to quantify the proportions of genetically distinct *P. falciparum* parasites occurring in mixed clone infections. These methods have already been reviewed in **Chapter 1**.

With the rapid technological advances, novel molecular techniques have been developed that are less costly, are more rapid and have higher throughput. Pyrosequencing™, a method developed by Ronaghi *et al*., 1996 was used in this study. Following Pyrosequencing™ technique validation for genotyping and allele quantitation in mixtures of *P. chabaudi chabaudi* rodent malaria parasites (Cheesman *et al*., 2006) the work outlined in this chapter adopted a similar strategy for analysis of *P. falciparum*.

3.1.1. Chapter synopsis

This chapter entails the SNP selection process from the available *P. falciparum* genomic databases. The Pyrosequencing™ technology for use in SNP analysis will also be clearly outlined. Finally, validation assays will be carried out using artificially prepared laboratory clones.

3.1.2. Aims of chapter

This chapter represents the technical aspects involved in assay design and the initial analysis using the selected SNPs and Pyrosequencing™. It therefore aims to:

1. Describe the Pyrosequencing™ technique in detail and justify its use in this study
2. Explain the process involved in SNP identification
3. Test the identified SNPs by Pyrosequencing™ using laboratory prepared pure *P. falciparum* clones.

3.1.3. Hypothesis

The testable hypothesis in this chapter states that the Pyrosequencing™ technique is capable of identifying and quantifying the SNPs obtained from the *P. falciparum* genome following the analysis of laboratory prepared pure clones of *P. falciparum*.

3.1.4. Materials and methods

The experiments performed in this chapter have been elaborately described in **Chapter 2**. The first step was to identify SNPs as detailed in the section below.

3.1.4.1. The SNP selection process

a) Selection of a chromosome arbitrarily from the *P. falciparum* genome and scanning it for conserved genes e.g. hypothetical genes. This is important to ensure that the selected SNPs were specific for the *P. falciparum* genome. This data was derived from the NCBI database (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=pfalciparum.inf).

b) The gene of interest was then retrieved from the PlasmoDB database (http://www.plasmodb.org/plasmo/home.jsp) and scanned for available SNPs. These SNPs were selected based on their ability to differentiate between the world wide *P. falciparum* field isolates whose sequences are available in the database and can be compared with the 3D7 reference genome. These SNPs were selected from loci that are:

   i. believed to be under no selection pressure. This is because alleles found in loci under high selection pressure may confound population genetic data interpretation as they will be highly represented in natural *P. falciparum* isolates

   ii. not highly polymorphic as these cause technical difficulties when designing primers for Pyrosequencing™ analysis

iii. not highly repetitive within the genome due to redundancy that may affect assay design and final analysis of results.

c) The genes containing SNPs of interest were retrieved from the databases and saved in a separate file. Bearing in mind that these genes were obtained from the 3D7 reference strain in the database, they were then aligned against the other available strains in the databases (HB3, Dd2 and 7G8 strains) using a sequence alignment tool derived from the website http://bioinformatics.unl.edu/emboss/. This was done for SNP verification with all the *P. falciparum* strains available in the databases.

d) Once the SNPs were verified, the sequences were saved in an MS Word file and stored for assay design.

Next, assay design was performed using the PSQ™ HS-96A assay design software as described in **Section 3.2.4**. The identified SNPs were then tested using laboratory prepared pure clones of *P. falciparum*. This was done to ensure that the developed assays and the technique used could provide the genetic identity of the tested parasite clones as well as quantitatively score them with accuracy giving the expected values of 100%.

3.2. Results

3.2.1. Pyrosequencing™ for genotyping and quantifying *P. falciparum* parasites

Pyrosequencing™ is a high throughput genotyping technique that involves the sequential incorporation of nucleotides on a growing strand of DNA hybridized on PCR pre-amplified template DNA. Its principle involves a series of enzyme-coupled reactions culminating in a light signal that is proportional to the number of nucleotides added to the sequence of DNA being synthesised. The enzyme cascade begins with the hybridisation of the sequencing primer onto the template DNA. Nucleotides are released sequentially and the nucleotide complementary to the template is added onto the synthesised DNA with the aid of DNA polymerase. Incorporation of the nucleotide produces inorganic pyrophosphate (PPi) which with the aid of the substrate adenosine 5′ phosphosulfate and the enzyme ATP sulfurylase produces ATP. The ATP is then converted to light in the presence of the substrate luciferin in a reaction catalysed by luciferase. The signal produced is detected by a charge-coupled device (CCD) camera and is visualised as a series of graphical presentations known as pyrograms (**Figure 3.1**). The unincorporated nucleotides and the excess ATP are degraded by the enzyme apyrase.

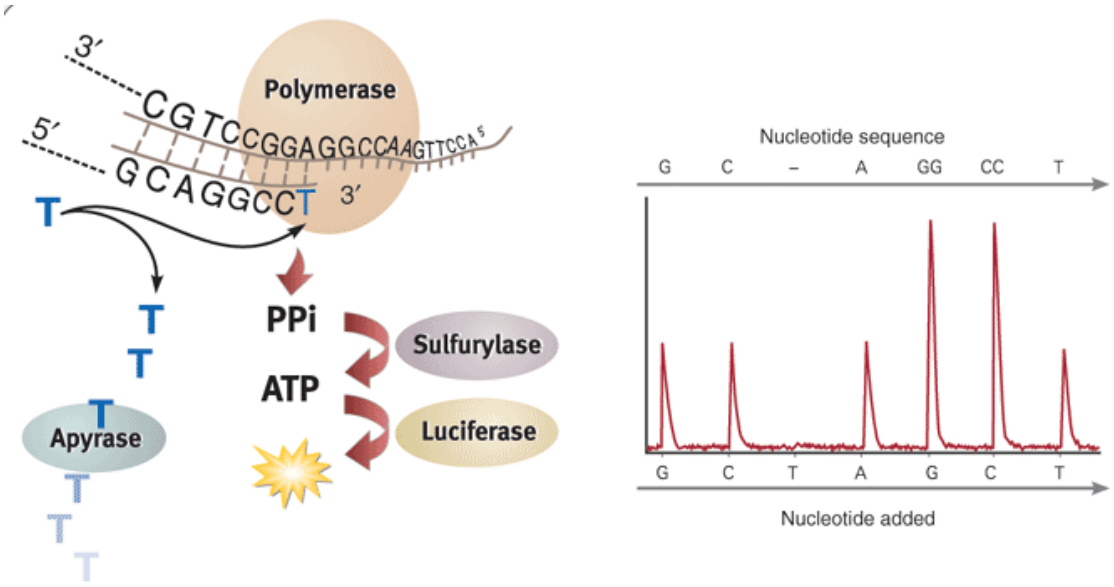


Figure 3.1 Diagrammatic presentation of Pyrosequencing™. The reaction is initiated by addition of an enzyme mixture containing DNA polymerase, ATP sulfurylase, luciferase and apyrase. The substrate mixture containing luciferin and adenosine 5′ phosphosulfate follows. The PCR product in the reaction consists of a single stranded DNA immobilised on a bead as described previously and serves as the template for the reaction. The sequencing primer anneals to the template and the reaction proceeds. In this example the template is the 3-CGTCCGGA…..-5′ strand shown above and the synthesised strand following nucleotide incorporation is 5′-GCAGGCCT…..-3′. The DNA polymerase binds to the region where the incoming nucleotide will hybridise as shown above. The incoming nucleotide (T) is complementary to the nucleotide on the template and is thus added to the growing strand with the aid of DNA polymerase. Once the T is added, inorganic pyrophosphate (PPi) is produced which in the presence of the substrate adenosine 5′ phosphosulfate and the enzyme sulfurylase leads to ATP production. The unused nucleotides (Ts) are degraded by the enzyme apyrase. The ATP is converted to light in the presence of the enzyme luciferase as shown above. This light is captured by a camera and is visualised in a computer as peaks known as pyrograms depicted on the right hand side of the figure above. The intensity of light produced and subsequently the height of the peak is dependent on the number of nucleotides added. For instance, as each nucleotide is sequentially added onto the strand being synthesised, a single peak is registered. When two nucleotides are added, this is registered as a double peak as shown above in the double peaks representing GG and CC Note that when the first T is added above, no peak is registered as it is not part of the analysed sequence. This is included in the reaction for quality control (Figure adapted from England & Pettersson, 2005).

This technique is most suited for this study as it genotypes and assigns quantitation values to parasites occurring in different proportions in a mixed clone infection as

shown in **Figure 3.2 a & b.** In addition, this method allows for programmed nucleotide dispensation order. If the sequence is known, only the nucleotides corresponding to the sequence under study are dispensed. The program therefore provides an instruction ensuring that unnecessary nucleotides are not dispensed unless they have been incorporated in the program for quality control or unless the nucleotides at the polymorphic region are unknown. This ensures high quality sequence reads (**Figure 3.2 a)** enabled by the high internal quality control in the instrumentation and subsequently in the final output.

(a)

```
TEMPLATE                3′      AT(T/C)CAGAT

NUCLEOTIDE SEQUENCE     5′      TA(A/G)GTCTA

NUCLEOTIDE DISPENSATION ORDER - GTACTGATCTA
```

(b)

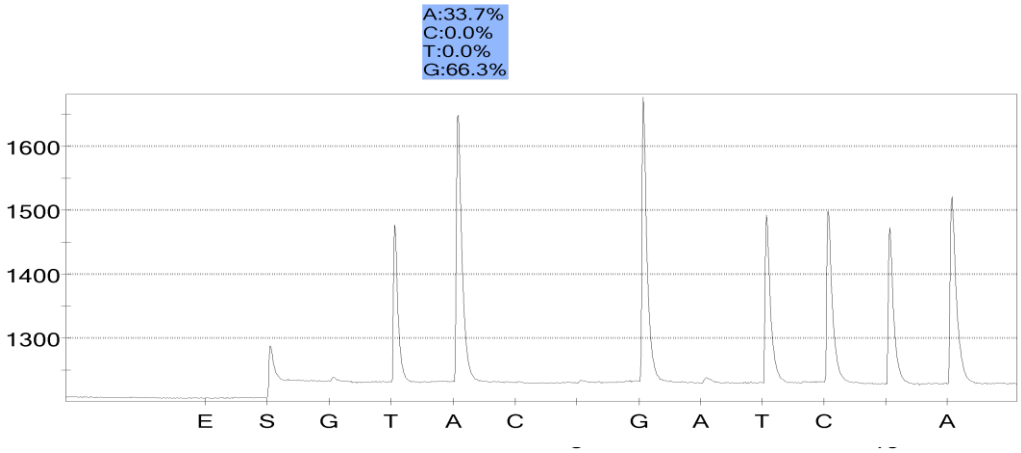


Figure 3.2 SNP analysis by Pyrosequencing™. This example is derived from analysis of a laboratory prepared artificial mixture of 3D7 and HB3 clones. **(a)** At the third (polymorphic) position in the sequence represented, 3D7 has the nucleotide T, and HB3 the nucleotide C, on the 3' (template) strand corresponding to A and G respectively on the read-out 5' strand. The Pyrosequencing™ pre-programmed nucleotide dispensation order is also shown. **(b)** The output of the Pyrosequencing™ analysis (pyrogram).

Pyrosequencing™ uses a pre-programmed nucleotide dispensation order following the complementary 5' sequence with the addition of the two nucleotides marked in red, G and A, in (a) above, as nucleotides included for quality control. The peaks in the pyrogram (b) above correspond to the nucleotides added onto the growing DNA strand. The E indicates addition of the enzyme mixture and S addition of the substrate mixture previously described in **Figure 3.1**.

The first G, as already stated, is included for (negative) quality control so no peak is registered. When a single T is then added (corresponding to the first A of the 3' sequence), a peak is registered. The next addition, an A, produces a larger peak

because the template has two successive T's (the complement of A) present in one of the strains as shown in the underlined region of the template (a) above. The C and T nucleotides are included in the dispensation order to cater for all the possible SNP combinations when analysing wild *P. falciparum* isolates. Since C and T are not represented in this example, no peak registers when they are released onto the growing strand. G is then added and a larger peak is again observed because two successive Cs are present in the template of one of the strains as underlined in (a) above. A negative control A is added, to which there is no signal, followed by four successive nucleotides - T,C,T and A - for four successive positions in which the sequence in both parasite strains is the same, each yielding a smaller sized peak corresponding to a single shared nucleotide.

The box highlighted in blue above the pyrogram in (b) gives the values assigned by the Pyrosequencing™ for the proportions of each nucleotide at the polymorphic (SNP) position and consequently of each clone in the mixture. Thus G, representing HB3, is calculated at 66.3% and A, representing 3D7, at 33.7%.

3.2.2. SNP identification

A single nucleotide polymorphism (SNP) is an alternative form of a nucleotide at a genetic locus (unique position in a DNA sequence) between members of the same species or, in this context, between two genetically distinct *P. falciparum* clones. SNPs can, therefore, be used as genetic markers to identify the various genetically distinct *P. falciparum* clones present in an infection. As earlier stated, *P. falciparum* mixed infections are common in nature and for this reason, SNPs as genetic markers allow the identification of parasite clones of different genotypes to be identified and distinguished in an infection.

3.2.3. Identifying SNP locations

The *P. falciparum* nuclear genome is ~23 Mb in size and is distributed into 14 chromosomes ranging in physical size from 640kb to 3290kb. It also contains a

separate plastid genome of approximately 35kb and a mitochondrial genome of 6kb (Gardner *et al*., 2002). The complete *P. falciparum* strain 3D7 genome sequence facilitated the sequencing and development of other databases containing the different *P. falciparum* clones (HB3, Dd2, 7G8, and so on). Various efforts have been made to develop a malaria haplotype map (Carlton, 2007). Of great significance and relevance to this study was the genome-wide SNP identification depicted in **Figure 3.3** by Mu *et al*., 2007. They re-sequenced Dd2, HB3, D10 and 7G8 clones and compared them with the 3D7 clone reference genome. This study revealed numerous genetic markers for use in genetic studies showing an average of one genetic marker per 3.6kb in the genome. A recent exploratory study to detect polymorphisms using a high-density microarray chip detected an average of one genetic marker per approximately 500bp in 3D7, DD2, 7G8, HB3 and FCR3 (Jiang *et al*., 2008). All the above studies highlight the abundance of SNPs across the *P. falciparum* genome available for use in this project.

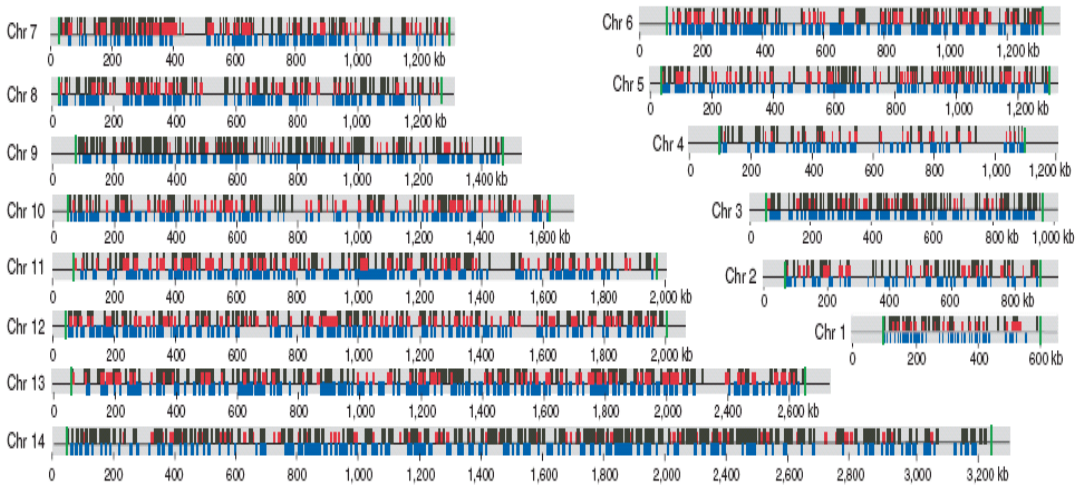


Figure 3.3 Distribution of SNPs across genome of *P. falciparum* derived from aligning cloned isolates, Dd2, HB3, D10 and 7G8 with the 3D7 genome sequence. The map details an average of one polymorphic marker per approximately 3.6 kb of the *P. falciparum* genome. Vertical bars represent SNPs (black, non-synonymous substitutions; red, synonymous substitutions) or microsatellites (blue, under the horizontal lines). Only one non-synonymous SNP and one synonymous SNP are presented if there was more than one SNP in a gene (non-coding SNPs are grouped with synonymous SNPs). Most of the chromosomal ends (green vertical bars) are excluded because of repetitive gene families such as *var*, *rifin* and *stevor* (From Mu *et al*., 2007).

The SNPs data can be accessed from the *P. falciparum* genome resource, PlasmoDB (http://www.plasmodb.org/plasmo/home.jsp) which provides information on polymorphisms occurring across the different cloned lines (3D7, HB3, Dd2 and so on) that have been sequenced and placed in the genome as representatives of worldwide isolates. Complete genomic data is usually first submitted to the PlasmoDB database then to the National Centre for Biotechnology Information (NCBI) database. Genomic data can be accessed with ease from both databases. These databases facilitated the selection of about 50 SNPs distributed in various chromosomes for use in this project.

3.2.4. Pyrosequencing™ assay design

Once genes of interest containing the SNPs best suited for identification of parasite clonal genotypes had been identified for Pyrosequencing™ analysis (as desbribed in **Section 3.1.4.1**), the next step was to design the specific assays for each of the SNPs. The Pyrosequencing™ instrumentation is furnished with assay design computer software. The assay design software allows for the uploading of files containing the genes of interest as long as the polymorphic regions are clearly highlighted. It then designs forward and reverse PCR primers for the amplification of DNA and an additional sequencing primer for the subsequent Pyrosequencing™ analysis. It then assigns scores for each of the assays it designs depending on how well-suited the gene region is for PCR and Pyrosequencing™ analysis. Bearing in mind that the *P. falciparum* genome is highly AT rich (>80%), (Gardner *et al*., 2002), good primer design can be difficult. Optimal primers must have a melting temperature (Tm) of 58-65°C which is usually determined by the GC content of the region you are designing primers from. The AT richness of the *P. falciparum* genome therefore renders primer design difficult. This computer software is highly pertinent as it designs assays around the polymorphic region and assigns assay scores giving warnings where:

i. The primer melting temperature (Tm) is too low which depends on GC content of the SNP region
ii. There is possibility of mis-priming due to non-specific binding of primer

iii. The template may generate secondary structures such as a hair pin loop structure, which compromise DNA polymerase progression along the template impeding further sequencing

Most of the SNPs residing in the pre-selected gene regions are, therefore, discarded at this point due to low assay scores (<70%) assigned by the computer software. It is on this basis that, of the 50 SNPs selected for use in this project from the PlasmoDB website, only 15 SNPs from four different chromosomes namely: Chromosome I, III, X & XIV (**Figure 3.4)** passed the Pyrosequencing^{TM} assay design software 'quality control' (see **Table 3.1**). The fact that these genetic markers were obtained from different chromosomes across the genome provided a good representation of sites for genotyping purposes.

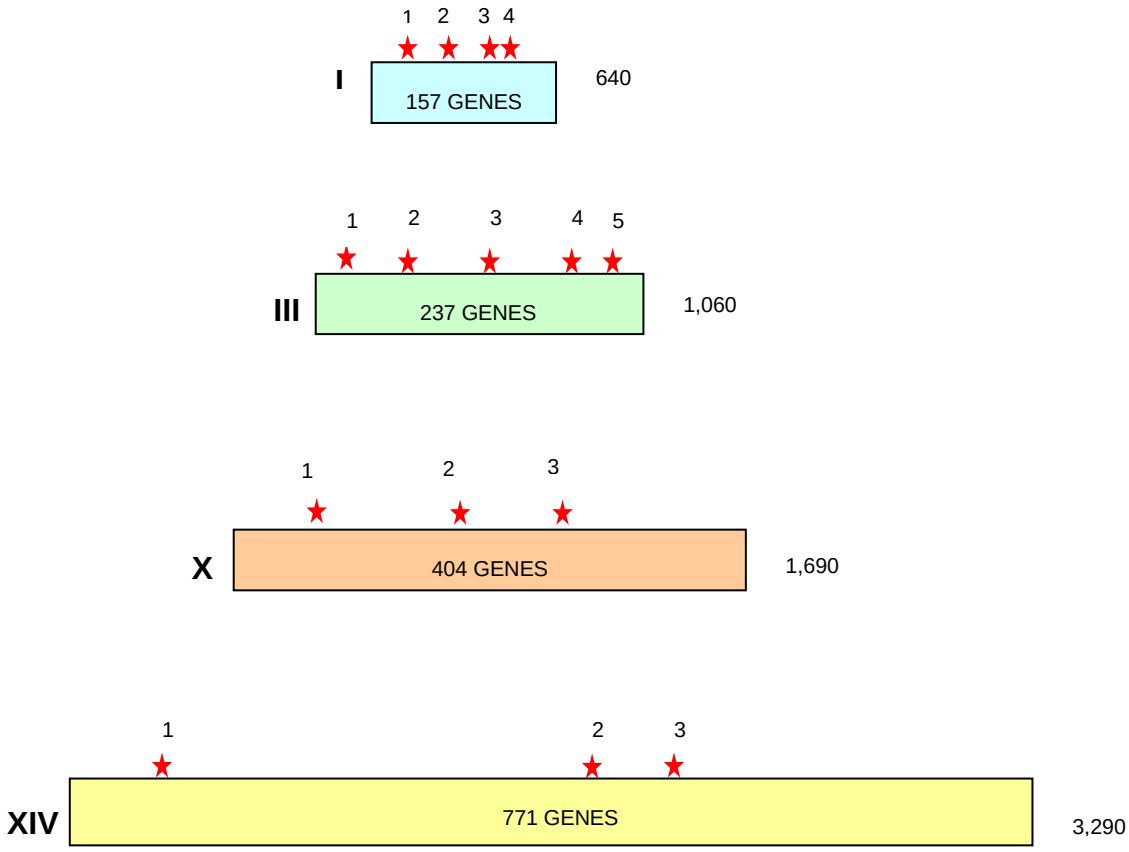


Figure 3.4 Diagram depicting the SNPs selected for genotype analysis shown by stars, and the genes in which they reside shown as red stars and the numbers 1-5. The genes are located in chromosomes I, III, X and XIV. The chromosomes are indicated by the coloured bars and the chromosome sizes in kilobases (kb) are given on the side of each bar. The total number of genes present in each chromosome is shown inside the chromosome bars. The genes are located in the approximate position they occupy in each chromosome. Note that the nomenclature varies in the PlasmoDB or NCBI website; the variant names are therefore provided.

Gene annotation:

Chromosome I: **Gene 1**. Ring infected erythrocyte surface antigen (PFA 0110w, PLASMODB or Mal1P1.13, NCBI); **Gene 2**. Hypothetical protein (PFA0230c, PLASMODB or Mal1P1.36, NCBI); **Gene 3**. Selenocysteine-specific elongation factor selB homologue, putative (MAL1P2.28); **Gene 4**. Phosphatidylinositol-4-phosphate-5-kinase, putative (MAL1P2.32).

Chromosome III: **Gene 1.** ABC transporter, putative (MAL 3P1.7); **Gene 2.** DNA polymerase δ small subunit, putative (MAL3P3.4); **Gene 3.** CTRP (PFC0640w); **Gene 4.** T-complex protein 1 ε subunit, putative (MAL3P7.9); **Gene 5.** Hypothetical protein (MAL3P7.37).

Chromosome X: **Gene 1.** Hypothetical protein (PF10_0031); **Gene 2.** 3.8 protein (Serine/Threonine protein kinase, FIKK family) (PF10_160); **Gene 3**. Hypothetical protein (PF10_230).

Chromosome XIV: **Gene 1.** Hypothetical protein (PF14_0074); **Gene 2.** Hypothetical protein, conserved (PF14_0400); **Gene 3.** Hypothetical protein (PF14_0485).

The assays designed for each of the SNPs by the Pyrosequencing™ assay design software and the assigned assay scores by the Pyrosequencing™ assay design software are shown in **Table 3.1**.

Table 3.1 The assays generated by the Pyrosequencing™ assay design software. The genes from where the SNPs occur and their sizes are given. The gene IDs derived from NCBI and PlasmoDB (in brackets) databases are given in the first column. The nomenclature of the allele specific primer sets given in the second column is: pf – *Plasmodium falciparum*; 01, 03, 10, and 14 denoting the chromosome from which they were derived; while the remaining digits indicate the physical position of the SNP in the genome. Since the genome was re-annotated, the SNP positions in some of the chromosomes changed; the new ones are given in brackets. The third column has the primer sets with the biotinylated ones given in bold while the fourth column shows the expected primer sizes following amplification. The last column shows the *quality control score (a 'quality index' (maximum 100) is allocated by the Pyrosequencing™ assay design software to each individual assay which is designed to reflect the quality of the assay).

Gene Annotation and size	Primer name	Pyrosequencing Primer sets (5′ – 3′). The biotinylated primers are in bold	Expected PCR Product size (bp)	*Quality control scores assigned by the Pyrosequencing™ assay design software
Ring-infected Erythrocyte surface antigen Mal 1P1.13 (PFA0110w) 3258 bp	Pf-01-101502	**F - GGATATTCAAACATGTCGCTAAAA** R – TGGTTTCCGTTATCAACATTGG S - CATTTGTTCACTACTTCCTA	214	90
Hypothetical protein Mal 1P1.36 (PFA0230c) 768 bp	Pf-01-205304	**F - TTATTTCCATTGTTTCGACTGTTG** R – AAGCAGTTTGATGAAAAGAACAA S - CCTGGATATGCAGATTT	184	79
Selenocysteine-specific elongation factor selB homologue, putative MAL 1P2.28 (PFA0495c) 2805 bp	Pf-01-393153	F – GGTGTTTTAGGACACGTTGATTCT **R – TTATGTTTATCTAGGGCGCATGT** S - CGAGCCTCTGTAAATGT	89	91
Phosphatidylinositol-4-phosphate 5-kinase, putative MAL1P2.32 (PFA0515w) 5133 bp	Pf-01-411796	**F – TTTTCATCCATTGTGGAGATACAT** R – CTCCCAATTGACCACATCTTTAGA S - AACGGTTCGTCCTACTA	282	77
ABC transporter, putative MAL3P1.7 (PFC0125w) 4098 bp	Pf-03-138-7 (Pf-03-139623)	F – AGAACGATCATTTCGTCTGTCATA **R – TTATTGTTGCGCATATTGTCCT** S - TGCAAAAATATAATTGCTCA	149	80
DNA polymerase δ small subunit, putative MAL3P3.4 (PFC0340w) 1497 bp	Pf-03-350224 (Pf-03-351177)	F - TGGGGATACATATTTATCAGAAAA **R – CTTTATCACCAGGCATTAAATCAA** S - GAAAAGGATATGAACACAAT	158	85
CTRP (PFC0640w) 6345 bp	Pf-03-613146 (Pf-03-614100)	**F – CAACACATTGCCAAGGTGAT** R – TTTCCTGCTCATAACCTATATCCG S - AACTACATCTTTCTTCCATT	222	80

Gene Annotation and size	Primer name	Pyrosequencing Primer sets (5′ – 3′). The biotinylated primers are in bold	Expected PCR Product size (bp)	*Quality control scores assigned by the Pyrosequencing™ assay design software
T-complex protein 1 ε subunit, putative MAL3P7.9 (PFC0900w) 1608 bp	Pf-03-842383 (Pf-03-842968)	F – ATATTGTTGTCGATGCCGTTTT **R – TTTAGGTGGTTCAAATGGACATG** S - TTATCAGTTGCAGATATGAA	218	81
Hypothetical protein MAL 3P7.37 (PFC1030w) 4629 bp	Pf-03-967838 (Pf-03-968192)	F – AGTGACACGGAATTATATGAGAAG **R – TTTGAAAACCTCCTTCAGAAAAA** S - TATATGAAAATGTGTCAACA	292	70
Hypothetical Protein PF10_0031 2835 bp	Pf-10-131053	**F – TGAAAAAGAAGTCATCGAAAGTAA** R – CTTTTGATTCTATTTTCGGTAGCA S - ATTTTTTCAACACTTTTATG	126	78
3.8 protein (Serine/Threonine protein kinase, FIKK family) (PF10_0160) 1857 bp	Pf-10-658467	**F – TGTGTAAACACGCTCCAATTAAAT** R – AATCGATGCTTAGCCGAGTTT S - TGCTTAGCCGAGTTTG	62	83
Hypothetical protein conserved (PF10_0230) 2274 bp	Pf-10-992433	F – TTGCCTATCATTAATGGAAAAT **R – TAATATAAATTCCGTACGGTCCTC** S - AAGATCTATCCAGTTCTTTA	165	76
Hypothetical protein (PF14_0074) 1221 bp	Pf-14-279668	F – ATTCTGAACAACGCCTTCCATTA **R – ACAGCAACAAAATCCTTCAGAGA** S - GTTTAGTTATTAGCCAGGTG	179	81
Hypothetical protein conserved (PF14_0400) 3003 bp	Pf-14-1710196	F – GGGAGCAAGTTTTTCTTCACTTTT **R – AAGTGATCATACCATTTTTGAAGC** S - AATGAACTTATTAAGTGTTT	262	71
Hypothetical Protein PF14_0485 2415 bp	Pf-14-2085411 (Pf-14-2086399)	F – GCTCAAACAATAGTAACAAAACCA **R – ATGTTGTGATGGTACATTCATGTC** S - AAAAACAAATGTTGAAAATT	222	78

The identified SNPs were then tested by Pyrosequencing™ using laboratory prepared pure clones of 3D7, HB3, CAMP and FCR3 to test the technique for SNP identification. The results shown in **Table 3.2** reveal that the technique does indeed identify each SNP representing the pure clones tested and corresponds with information provided in the PlasmoDB database. The proportions assigned by the Pyrosequencing™ are also shown with all the SNPs occurring at a proportion >95% out of the expected 100%.

Table 3.2 Basic operational information on the SNPs used for genotyping. The primer names, SNP positions in the genes where the assays are derived from are given and the genotypes for the sequenced *P. falciparum* clones available in the PLASMODB database. The table shows that the SNP determined by Pyrosequencing™ analysis in each locus corresponds to that provided in the database as shown in the observed *P. falciparum* strain and SNP type column. This data was obtained from pure clones of 3D7, HB3, FCR3 and CAMP. The expected proportions assigned by the Pyrosequencing™ should be 100%. As shown below, the observed mean allele quantification values obtained from Pyrosequencing™ are all very close to100%.

Primer name	**SNP position in gene and genotype of sequenced clones from the database**	**SNP type**	**Observed *P. falciparum* strain and SNP type**	**Mean proportion of SNPs (%) assigned by Pyrosequencing™ from the analysis of pure clones**
Pf-01-101502	2245 3D7/RO33 - **G** 7G8/D10/D6/DD2/FCR3/K1/HB3/IT/SENEGAL3404 – **T**	Non-synonymous	3D7 - G HB3 – T FCR3 - T	98.6
Pf-01-205304	324 3D7/DD2/7G8/D10/RO33/SANTALUCIA – **G** HB3 -**A**	Synonymous	3D7 - G HB3 - A	99.1
Pf-01-393153	67 Dd2 – **T** 3D7/7G8/FCC-2/FCR3/GHANA1/HB3- **C**	Synonymous	3D7 – C CAMP – C FCR3 – C HB3 - C	98.8
Pf-01-411796	4618 DD2/IT - **G** 3D7/D10/ /GHANA1/HB3/VS-1 - **T**	Non-synonymous	3D7 - T CAMP – T FCR3 – G HB3 – T	98
Pf-03-138-7	1389 3D7/ HB3 – **G** DD2/D10/7G8 - **T**	Synonymous	3D7 - G CAMP - T FCR3 – G HB3 - G	97.3

Primer name	SNP position in gene and genotype of sequenced clones from the database	SNP type	Observed *P. falciparum* strain and SNP type	Mean proportion of SNPs (%) assigned by Pyrosequencing™ from the analysis of pure clones
Pf-03-350224	954 3D7 – **C** DD2/7G8/D10/HB3 - **T**	Synonymous	3D7 - C CAMP - T FCR3 – C HB3 - T	98.6
Pf-03-613146	955 3D7 – **G** DD2/HB3/7G8/D10 - **A**	Non-synonymous	3D7 – G CAMP - G FCR3 – A HB3 - A	98.5
Pf-03-842383	586 HB3 – **C** 3D7/DD2/D10/7G8 - **A**	Synonymous	3D7 - A CAMP - A FCR3 – A HB3 - C	96.5
Pf-03-967838	355 3D7 – **C** D10/DD2/HB3/7G8 - **A**	Non-synonymous	3D7 - C CAMP - A FCR3 – A HB3 - A	99.2
Pf-10-131053	1777 7G8- **T** 3D7/DD2/FCC2/GHANA1/HB3/ SENEGAL 34.04/V1-S - **C**	Synonymous	3D7 - C CAMP - C FCR3 – C HB3 - C	95.4
Pf-10-658467	278 3D7 - **C** 7G8/FCR3/D6/HB3/SANTA LUCIA/D10/K1/FCC-2 – **T**	Non-synonymous	3D7 - C HB3 – T	100
Pf-10-992433	1289 3D7/7G8/RO33 – **A** HB3/SANTALUCIA/GHANA1 – **G** 106_1/D10/DD2/FCC-2/K1/V1_S - **C**	Non-synonymous	3D7 - A HB3 – G	100
Pf-14-279668	493 3D7/RO33/SENEGAL3404 – **A** HB3/7G8/D6/DD2/FCB/K1/FCC-2/V1_S/IT - **T**	Non-synonymous	3D7 - A HB3 – T	99.7

Primer name	SNP position in gene and genotype of sequenced clones from the database	SNP type	Observed *P. falciparum* strain and SNP type	Mean proportion of SNPs (%) assigned by Pyrosequencing™ from the analysis of pure clones
Pf-14-1710196	1278 3D7/7G8 – **A** HB3/D10/D6/DD2/K1/RO33/SENEGAL3404/V1_S/K1 - **T**	Synonymous	3D7 - A HB3 – T	97.7
Pf-14-1710196	1278 3D7/7G8 – **A** HB3/D10/D6/DD2/K1/RO33/SENEGAL3404/V1_S/K1 - **T**	Synonymous	3D7 - A HB3 – T	97.7
Pf-14-2085411	1058 GHANA1/VS-1/Dd2/SENEGAL34.04/HB3/D10/3D7 – **T** 7G8 - **G**	Non-synonymous	3D7 - T CAMP - T FCR3 – T HB3 - T	95.7

3.3. Discussion

In this chapter, (i) the SNPs of interest have been identified, (ii) assays for SNP analysis have been designed and (iii) the assays have been tested using pure clones of *P. falciparum* with the results ascertaining their use by Pyrosequencing™. The results obtained in this thesis conform to the hypothesis that indeed the technique identified and assigned the expected quantitation values (100%) to the analysed pure *P. falciparum* clones.

At the writing of this thesis, tremendous advances had been made on the identification and use of genome wide SNPs for *P. falciparum* clonal identification as discussed in **Chapter 1**. In addition, the PlasmoDB website has been constantly

updated with studies conducted on worldwide isolates increasing the SNP data. For example Nkhoma *et al*., 2012 used a panel of 384 SNPs obtained from the PlasmoDB database. These SNPs were selected on the basis of being highly polymorphic in Africa enabling detection of the extensive mixed clone parasite repertoire typically observed in this high malaria transmission region.

The availability of SNP data as described by Nkhoma *et al*., 2012 and with more SNPs being identified and tested at the Sanger Institute such as by Campino *et al*., 2011 or Manske *et al*., 2012 eases the burden of scouring through the PlasmoDB or NCBI websites as was done in this study. Pyrosequencing™ assays can therefore be designed around these SNPs. Furthermore, as observed in this chapter, the Pyrosequencing™ assay design software is highly stringent such that of the 50 identified genes of interest for SNP detection only 15 were retained. Thus with the resource of a large number of SNPs more assays can be designed for use with this technique.

A major setback observed so far with this technique is the rigorous and time-consuming process of assay design and testing. However, the advantage of this technique is that once an assay has been ascertained for use, the interpretation of data obtained from the pyrogram reads is straightforward and does not need extensive downstream proofreading as is required with other genome wide sequencing techniques. For example the deep sequencing analysis performed using Illumina platforms such as Solexa requires extensive sequence filtering *in silico* to ensure only quality and genuine reads are retained.

Further to this, Pyrosequencing™ does not require the removal of human DNA by a leukocyte depletion step as is recommended when analysing *P. falciparum* field isolates using the Illumina platforms (Auburn *et al*., 2011). This is because in the PCR step, primers specific for the amplification of parasite DNA are used.

3.3.1. Conclusion

In this chapter, 15 SNPs have been selected and their assays developed for analysis by Pyrosequencing™. More SNPs can indeed be identified for use especially with the ever increasing SNP data from worldwide *P. falciparum* isolates in the PlasmoDB database. Since the work outlined in this thesis constitutes a feasibility study, the identified SNPs should suffice in this preliminary analysis. The SNPs will then be validated for use in the next chapter using laboratory prepared mixed clone infections as well as material from the cross progeny obtained from the 3D7 and HB3 parasite clones.

4. STANDARDIZATION OF PYROSEQUENCING™ ASSAYS

4.1. Introduction

P. falciparum mixed clone infections are common in nature and parasites in such infections are usually present at different proportions. This is a situation that can be exploited and, with a technique that can identify SNPs representing the individual parasite lines in a mixed infection and assigning proportions to them, it is possible to classify individual alleles existing at similar proportions as belonging to the same genotype. These sets of alleles representing individual parasite clones in an infection, can be used to trace genetic relatedness of parasites where, the more closely related the parasites are, the larger the proportion of alleles they share.

4.1.1. Synopsis of chapter

Pyrosequencing™ has been tested using laboratory prepared pure clones of *P. falciparum* and is validated for use in this study (see **Chapter 3**). The next objective was to determine if Pyrosequencing™ is able to quantify proportions of parasites carrying individual SNPs in a mixed clone infection with sufficient accuracy to identify parasite genotypes represented in laboratory prepared parasite mixtures. This is in line with work previously outlined by Cheesman *et al*., (2007).

Also in this chapter, prior to analysis of *P. falciparum* field isolates, the genetic relatedness in a cross comprising known parental clones (3D7 and HB3) will be

determined. Unlike the previous analysis involving laboratory prepared clones at known proportions, the cross progeny comprises genetically distinct clones, 3D7 and HB3 at unknown proportions as is commonly observed in field isolates. In this chapter, the accurate detection and delineation of the 3D7 or HB3 clonal genotypes occurring within the cross will be tested in preparation for analysis of field isolates.

4.1.2. Aims of chapter

The experiments in this Chapter were, therefore, conducted to

1. Determine if Pyrosequencing™ is able to accurately determine the genetic identity of parasites in mixtures of *P. falciparum* blood stage parasites of different genotypes.
2. Identify the technique's limits of detection in isolates occurring at different parasitaemia levels and how this may confound detection of parasite clones.
3. Identify clonal genotypes represented in the cross progeny obtained from the genetically distinct clones, 3D7 and HB3.
4. Test the genetic relatedness observed from analysis of the cross progeny.

4.1.3. Hypothesis

The testable hypotheses in this chapter are that:

1. Clonal genotypes can be accurately discriminated from the laboratory prepared parasite mixtures using the identified SNPs by Pyrosequencing^{TM}.

2. The *P. falciparum* parental clones, 3D7 and HB3 are expected to mate randomly such that the progeny comprises 50% parental genotypes and 50% recombinants.

4.1.4. Materials and methods

4.1.4.1. *P. falciparum* clonal mixtures

Laboratory prepared *P. falciparum* parasite mixtures containing two different genetically distinct clones, namely 3D7 & HB3, 3D7 & DD2 or 3D7 & 7G8 were made in various proportions to test this (**See Chapter 2**). These mixtures were prepared thus to mimic mixed parasite genotype infections that, as already stated, are commonly found in natural *P. falciparum* infections.

The mixed clone combinations were prepared at 0.1, 0.01 & 0.001% final parasitaemias and in calculated proportions of 100:0, 98:2, 95:5, 90:10, 80:20, 50:50, 20:80, 10:90, 5:95, 2:98 and 0:100 (made by Richard Carter). All the Pyrosequencing™ SNP assays previously generated for use in this study (**Chapter 3**) were tested. For each of these SNP assays, Pyrosequencing™ was tested for its ability to assign the proportions of the parasites of each genotype in a mixture. Duplicate samples of each combination of calculated parasite proportions were PCR amplified and triplicate Pyrosequencing™ measurements of each were taken.

4.1.4.2. Analysis of cross progeny

Information on how the *P. falciparum* cross progeny was cultured and semi-cloned is given in **Chapter 2**. Following identification of the clonal genotypes, estimates of genetic distances (GDs) and genetic relatedness (GR) were determined from overall pair-wise allelic comparisons for all loci analysed in all the recombinants (from **Figure 5.2**) using the formula:

$$GD = 1 - \frac{\text{No. of alleles shared}}{\text{Total alleles analysed}}$$

$$GR = 1 - GD$$

4.2. Results

4.2.1. Analysis of mixtures of genetically distinct clone by Pyrosequencing™ to determine the genotypes of each of the clones represented in the mixtures

The mean values derived from the medians of each triplicate measurement of SNP proportions representing each of the genetically distinct parasites in a mixture are summarised in **Table 4.1 a, b & c**.

Table 4.1 Summary of results obtained from Pyrosequencing™ analysis of laboratory prepared parasite mixtures at 0.1% parasitaemia of different proportions of (a) 3D7 & HB3, (b) 3D7 & DD2 and, (c) 3D7 & 7G8. Calculated values for the proportions in which the parasite mixtures were prepared are shown in the left hand column. The measured proportions of SNPs at each marker locus are those assigned by Pyrosequencing™ for each SNP assay, as indicated, e.g. pf-01-101502 etc.

(a)

Calculated values for each clone	Measured proportions of SNPS at each marker locus							
	pf-01-101502	pf-01-205304	pf-03-350224	pf-03-842383	pf-03-967838	pf-10-658467	pf-10-992433	pf-14-1710196
3D7 (100)	89.6	99.1	99.1	100	93.1	96.7	99.3	100
HB3 (0)	10.4	0.9	0.9	0	6.9	3.3	0.7	0
3D7 (98)	90.6	96	96.1	95	90.5	95.3	96.4	94.8
HB3 (2)	9.4	4	3.9	5	9.5	4.7	3.6	5.2
3D7 (95)	88.1	95.5	87.8	92.6	86.2	94.5	92.7	92.3
HB3 (5)	11.9	4.5	12.2	7.4	13.8	5.5	7.3	7.7
3D7 (90)	78.4	87.6	80.4	82.7	80.1	90.6	88.7	85.2
HB3 (10)	21.6	12.4	19.6	17.3	19.9	9.4	11.3	14.8
3D7 (80)	73.9	77.9	73.5	74.9	69.5	84	76.4	75.5
HB3 (20)	26.1	22.1	26.5	25.1	30.5	16	23.6	24.5
3D7 (50)	57.3	66.7	61.8	63	62.9	65.5	67.4	61.1
HB3 (50)	42.7	33.3	38.2	37	37.1	34.5	32.6	38.9
3D7 (20)	45.5	46.4	51.2	59.2	48.9	47.4	48.1	49.4
HB3 (80)	54.4	53.6	48.8	40.8	51.1	52.6	51.9	50.6
3D7 (10)	26.5	34.6	33.2	35.7	33.2	39.6	33.9	31.4
HB3 (90)	73.5	65.4	66.8	64.3	66.8	60.4	66.1	68.6
3D7 (5)	8.9	18.6	18.5	21.5	19.5	31.9	22.8	16.1
HB3 (95)	91.1	81.4	81.5	78.5	80.5	68.1	77.2	82.9
3D7 (2)	12	12.9	17	19.7	14.9	20.6	16.2	11.2
HB3 (98)	88	87.1	83	80.3	85.1	79.4	83.8	88.8
3D7 (0)	0.4	7.4	7.4	11.2	15.1	17.3	6.8	0.9
HB3 (100)	99.6	92.6	92.6	88.8	84.9	82.7	93.2	99.1

(b)

Calculated values for each clone	Measured Proportions of SNPs at each marker locus						
	pf-01-393153	pf-01-411796	pf-03-350224	pf-03-138-7	pf-03-967838	pf-14-1710196	pf-14-279668
3D7 (100)	100	100	94.7	95.6	85.5	88.7	85.8
DD2 (0)	0	0	5.3	4.4	14.5	11.3	14.2
3D7 (98)	98.1	95.6	91.4	94.6	87.7	91	85.6
DD2 (2)	1.9	4.4	8.6	5.4	12.3	9	14.4
3D7 (95)	95.3	95.4	89.5	90.4	88.9	89.2	83.8
DD2 (5)	4.7	4.6	10.5	9.6	11.1	10.8	16.2
3D7 (90)	93.1	88.5	86.0	87.2	82.7	82.6	83
DD2 (10)	6.9	11.5	14.0	12.8	17.3	17.4	17
3D7 (80)	83.1	79.8	72.1	84.2	75.1	72.9	78
DD2 (20)	16.9	20.2	27.9	15.8	24.9	27.1	22
3D7 (50)	56.5	48.9	46.9	47.7	53.5	54.5	55.5
DD2 (50)	43.5	51.1	53.1	52.3	46.5	45.5	44.5
3D7 (20)	29.6	25.2	17.9	22.3	22.6	21.9	36.6
DD2 (80)	70.4	74.8	82.1	77.7	77.4	78.1	63.4
3D7 (10)	19.5	14.7	7.8	13.8	17.5	16.4	23
DD2(90)	80.5	85.3	92.2	86.2	82.5	83.6	77
3D7 (5)	16.2	14.2	2.2	10.7	0	11	28.1
DD2 (95)	83.8	86.8	97.8	89.3	100	89	71.9
3D7 (2)	14.4	13.2	2.7	8.3	0	14	24.8
DD2 (98)	85.6	86.8	97.3	91.7	100	86	75.2
3D7 (0)	12.5	8.3	1.8	2.5	0	6.5	25.7
DD2 (100)	87.5	91.7	98.2	97.5	100	93.5	74.3

(c)

Calculated values for each clone	Measured proportions of SNPs at each marker locus					
	pf-03-138-7	**pf-03-613146**	**pf-03-350224**	**pf-10-131053**	**pf-10-658467**	**pf-14-2085411**
3D7 (100)	98.9	94	96.7	89.5	92.6	98.7
7G8 (0)	1.1	6	3.3	10.5	7.4	1.3
3D7 (98)	96.8	93.7	93.3	93.2	92	97.3
7G8 (2)	3.2	6.3	6.7	16.8	8	2.7
3D7 (95)	96.9	94.9	94	89.2	89.3	96.3
7G8 (5)	3.1	5.1	6	10.8	10.7	3.7
3D7 (90)	93.6	91.6	93.8	92	91.3	96.8
7G8 (10)	6.4	8.4	6.2	8	8.7	3.2
3D7 (80)	91.4	89	88.5	86.7	89.1	94.4
7G8 (20)	8.6	11	11.5	13.3	10.9	5.6
3D7 (50)	70.3	70.1	69.9	67.1	60.3	75.6
7G8 (50)	29.7	29.9	30.1	32.9	39.7	24.4
3D7 (20)	46.6	40.4	40.5	48.2	49.3	54.1
7G8 (80)	53.4	59.6	59.5	51.8	50.7	45.9
3D7 (10)	44.1	21	36.4	17.9	19.9	25.3
7G8(90)	55.9	79	63.6	82.1	80.1	74.7
3D7 (5)	23.2	14.7	22.4	4.8	9.8	20.0
7G8 (95)	76.8	85.3	77.6	95.2	90.2	80.0
3D7 (2)	18.9	4.8	15.1	0.1	0.7	10.0
7G8 (98)	81.1	95.2	84.9	99.9	99.3	90.0
3D7 (0)	18.4	12	8.7	1.3	0.4	1.0
7G8 (100)	81.6	88	91.3	98.7	99.6	99.0

For each marker locus, Pyrosequencing™ quantifies the SNP for each parasite strain in a mixture. This, therefore, indicates the proportion of parasites represented by that SNP in the mixture (**Table 4.1 a, b & c).** In principle, all SNPs which are represented in the same proportion must belong to parasites of the same genotype.

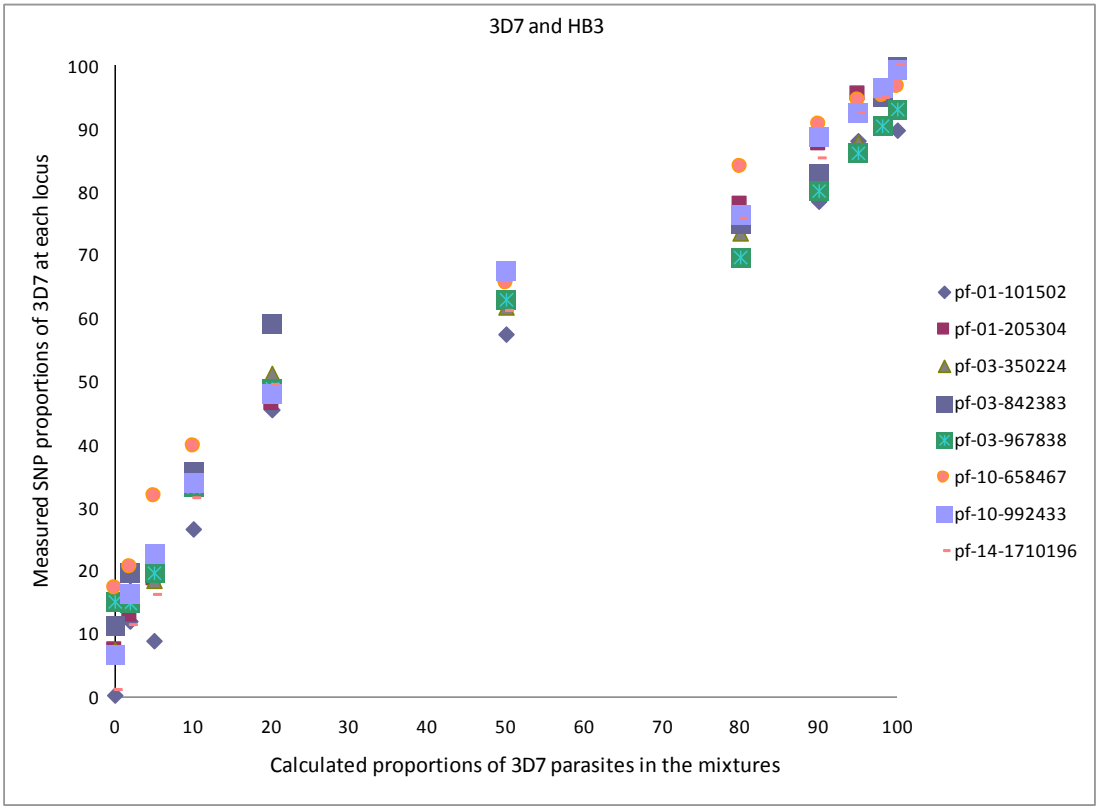
The results in **Table 4.1 a, b & c** showed that, when a clone was represented in high proportion in a mixture (the major clone), its SNPs were measured at the same high proportion and there was usually no difficulty in identifying its genotype. In this way, for most parasite strain combinations, correct strain genotypes could be determined at least for the major clone in the mixture.

However, Pyrosequencing™ has limited accuracy to assign SNP proportions for clones occurring in low proportions (**Table 4.1 a, b & c)**. In general Pyrosequencing™ is inaccurate in assigning measurements for SNP proportions for clones occurring in proportions of <25%. Parasites and their SNPs that were represented in apparent proportions of less than 25% could not be confidently differentiated from total absence, i.e. from 0%. It is also not possible to clearly distinguish SNP proportions for clones occurring at proportions between 40% and 60%. Conversely, for major clones represented at proportions >60% or with minor clones represented between 25-40%, a SNP genotype could normally be confidently given. On this basis, it can be concluded that Pyrosequencing™ quantifies, within certain limits of accuracy, the proportion of a SNP at each marker locus for each parasite genotype in a mixture.

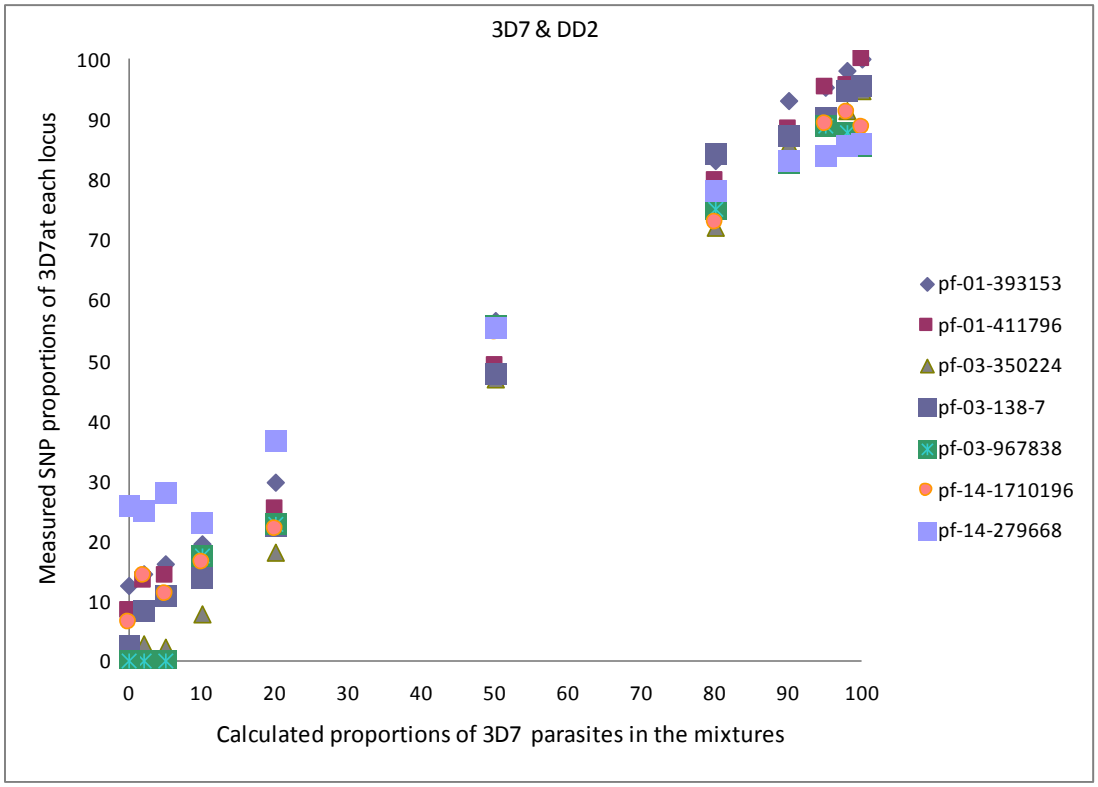
4.3. Accuracy and precision of Pyrosequencing™ in assigning proportions relative to the calculated values of the laboratory prepared genetically distinct clones

To test the reliability of Pyrosequencing™ in assigning SNP proportions relative to the calculated values of the clones represented in the laboratory prepared mixtures, the results in **Table 4.1 a, b & c** were plotted in a graph (see, **Figure 4.1 a, b & c)**.

(a)



(b)


3D7 & DD2
Measured SNP proportions of 3D7at each locus
Calculated proportions of 3D7 parasites in the mixtures
0
10
20
30
40
50
60
70
80
90
100
pf-01-393153
pf-01-411796
pf-03-350224
pf-03-138-7
pf-03-967838
pf-14-1710196
pf-14-279668

c)

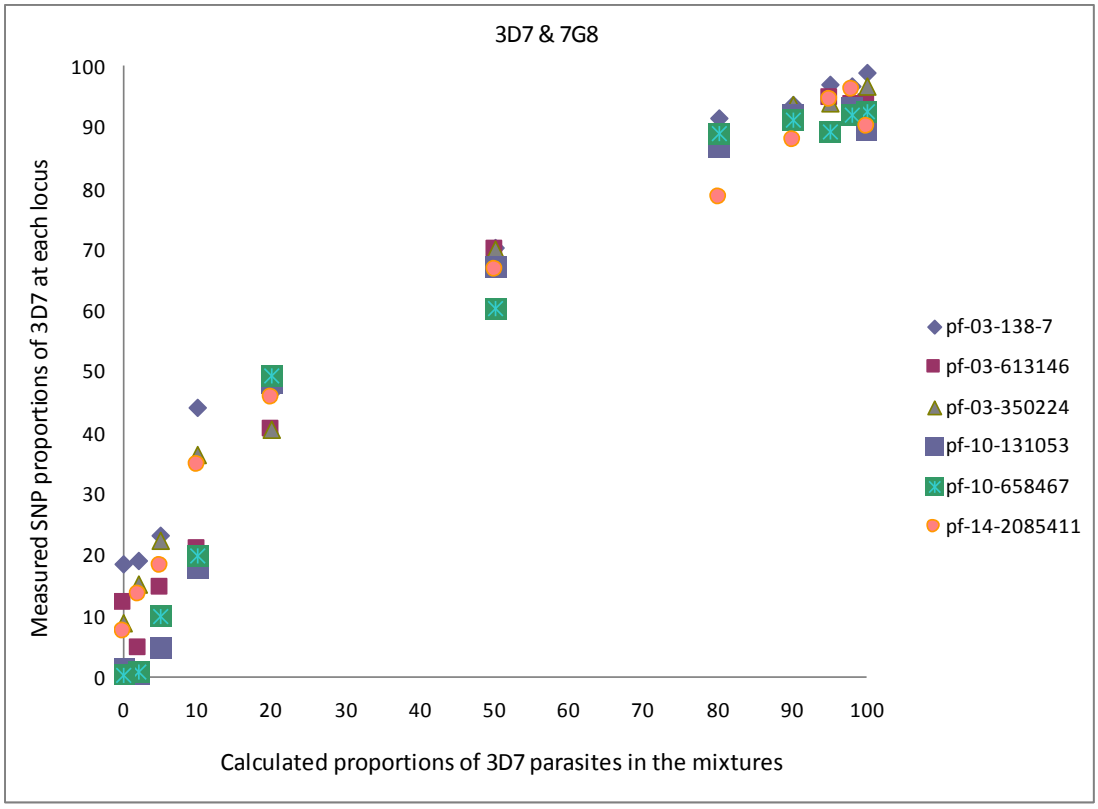
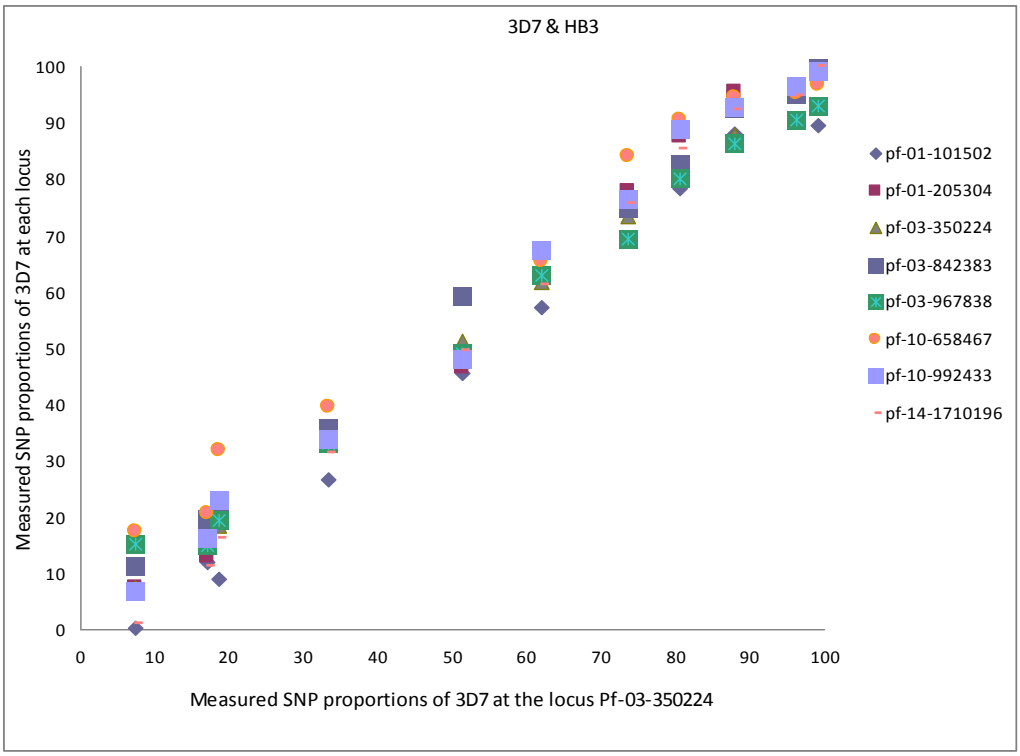

3D7 & 7G8
Measured SNP proportions of 3D7 at each locus
Calculated proportions of 3D7 parasites in the mixtures
0
10
20
30
40
50
60
70
80
90
100
pf-03-138-7
pf-03-613146
pf-03-350224
pf-10-131053
pf-10-658467
pf-14-2085411

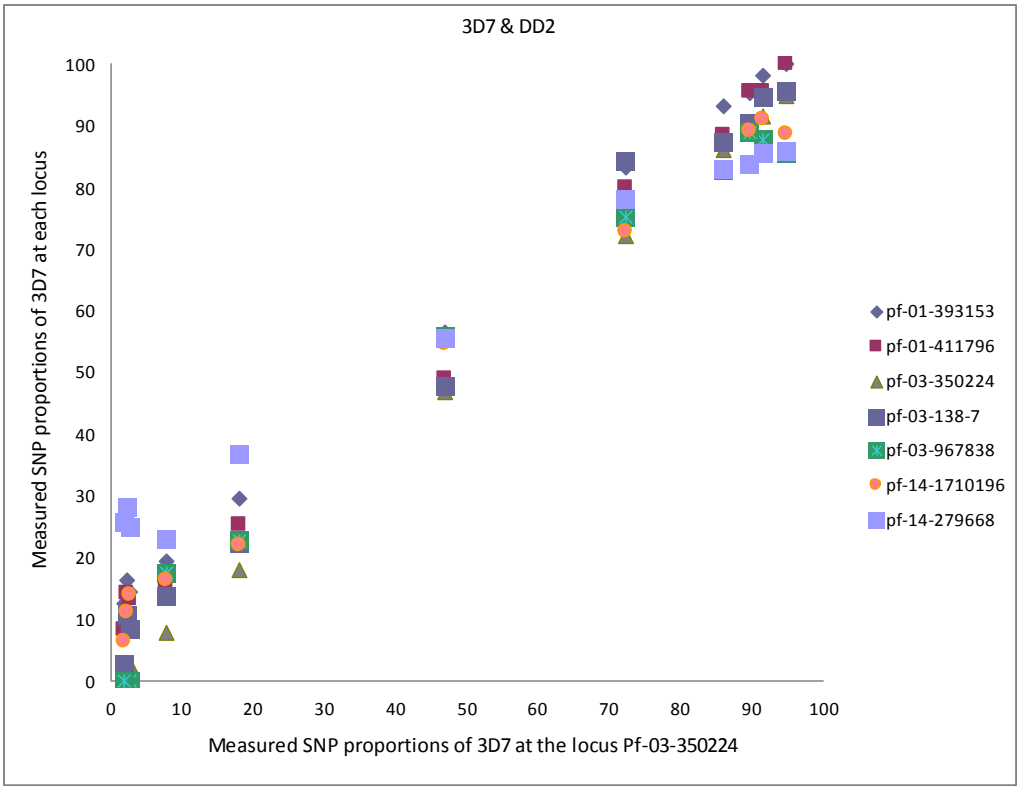
Figure 4.1 Diagrammatic presentation of the data in **Table 4.1 a, b & c**. The Pyrosequencing™ measurements of SNPs at all the analysed marker loci were plotted against calculated proportions of the parasites of each genotype in the laboratory prepared parasite mixtures of 3D7 & HB3, 3D7 & DD2 and, 3D7 & 7G8.

The results in **Figure 4.1 a, b & c** reveal a nonlinear relationship between the calculated values and the Pyrosequencing™ measurements. For each combination of parasite genotypes, the SNP proportions assigned by Pyrosequencing™ all showed closely similar deviations from the calculated proportions of parasite genotypes in the mixture (see, **Figure 4.1 a, b & c)**. The patterns, or directions, of these deviations were unique to each parasite mixture analysed (**Figure 4.1 a, b & c**). It is highly probable that these results are accounted for by inaccuracies in the calculated proportions of the parasites in the preparation of the laboratory mixtures. This is indicated by the fact that the discrepancies were unique to each mixture but closely consistent among the measurements of the different Pyrosequencing™ assays for the same mixture. To test the degree of internal consistency among the measurements assigned by the different Pyrosequencing™ assays for the same parasite preparation, the SNP proportions measured by Pyrosequencing™ at all loci were plotted against the SNP proportions measured at the locus Pf-03-350224. This was done for each of the laboratory prepared parasite mixtures (**Figure 4.2 a, b & c**).

(a)


3D7 & HB3
Measured SNP proportions of 3D7 at each locus
Measured SNP proportions of 3D7 at the locus Pf-03-350224
pf-01-101502
pf-01-205304
pf-03-350224
pf-03-842383
pf-03-967838
pf-10-658467
pf-10-992433
pf-14-1710196

(b)


3D7 & DD2
Measured SNP proportions of 3D7 at each locus
Measured SNP proportions of 3D7 at the locus Pf-03-350224
pf-01-393153
pf-01-411796
pf-03-350224
pf-03-138-7
pf-03-967838
pf-14-1710196
pf-14-279668

(c)

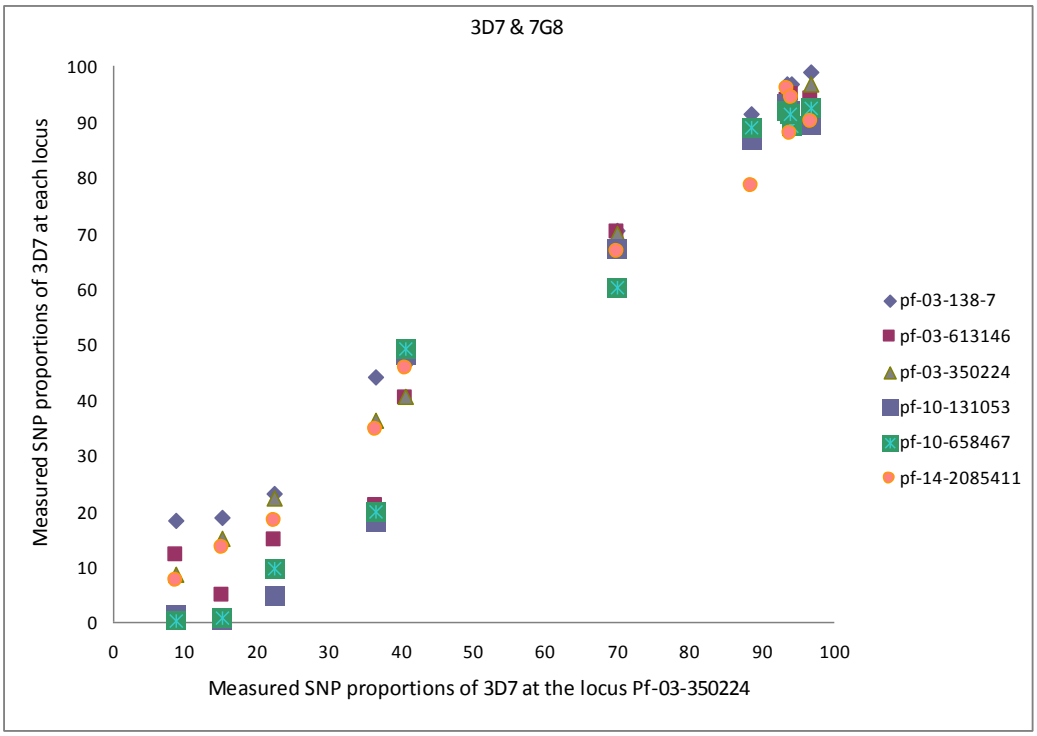


Figure 4.2 SNP proportions measured by Pyrosequencing™ at all the marker loci plotted against measured SNP proportions at the locus Pf-03-350224 for each of the parasite mixtures derived from analysis of 3D7 & HB3, 3D7 & DD2, and, 3D7 & 7G8 at 0.1 % parasitaemia

The results in **Figure 4.2 a, b & c** demonstrate relative consistency among all the Pyrosequencing™ assays in the SNP proportions assigned for each parasite mixture. This shows that, indeed, the deviations observed in **Figure 4.1 a, b & c** might have arisen due to relative inconsistencies during preparation of the laboratory prepared mixtures thus the proportions assigned by Pyrosequencing™ are accurate.

These results, therefore, indicate that Pyrosequencing™ can, within the limits identified, be reliably used to measure the proportions of SNPs representing parasites of different genotypes in many, but not all, mixtures containing genetically

distinct parasites. From this information it is usually possible to assign at least one and often two of the parasite genotypes present in a mixture. On this basis it should be possible to apply this method to assign genotypes in *P. falciparum* field isolates containing mixtures of genetically distinct parasites.

4.4. Testing the limit of detection of Pyrosequencing™ from P. falciparum genetically distinct clone mixtures prepared at 0.01% and 0.001% parasitaemias

To validate the use of Pyrosequencing™ for analysis of field isolates whose parasitaemias vary, it was pertinent to test samples with reduced parasitaemias to determine the technique's limit of detection i.e. at what level of parasitaemia it would still be possible to identify parasite genotypes from mixtures. *P. falciparum* laboratory prepared clone mixtures containing 3D7 & HB3, 3D7 & DD2 or 3D7 & 7G8 at proportions previously described and at parasitaemias of 0.01% and 0.001% were also tested by Pyrosequencing™. The markers summarised in **Table 4.2** were used in this analysis and their results are given in **Table 4.3 (a, b & c)**.

Table 4.2 Summary of results obtained from Pyrosequencing™ analysis of laboratory prepared parasite mixtures at 0.1%, 0.01% and 0.001% parasitaemia of different proportions of (a) 3D7 & HB3, (b) 3D7 & DD2 and, (c) 3D7 & 7G8. Calculated values for the proportions in which the parasite mixtures were prepared are shown in the left hand column. The measured proportions of SNPs at each marker locus are those assigned by Pyrosequencing™ for each SNP assay, as indicated, e.g. pf-01-205304 etc.

(a)

Calculated values for each clone	Measured Proportions of SNPs at each marker locus														
	pf-01-205304			pf-03-350224			pf-03-842383			pf-10-658467			pf-10-992433		
	0.10	0.010	0.001	0.10	0.010	0.001	0.10	0.010	0.001	0.10	0.010	0.001	0.10	0.010	0.001
3D7 (100)	99.1	100.0	92.5	99.1	89.8	81.1	100	100.0	81.2	96.7	94.0	77.0	99.3	88.9	78.1
HB3 (0)	0.9	0.0	7.5	0.9	10.2	18.9	0	0.0	18.8	3.3	6.0	23.0	0.7	11.1	21.9
3D7 (98)	96	98.8	83.9	96.1	81.5	78	95	96.1	71.5	95.3	85.7	78.3	96.4	91.6	75.1
HB3 (2)	4	1.2	16.1	3.9	18.5	22	5	3.9	28.5	4.7	14.3	21.7	3.6	8.4	24.9
3D7 (95)	95.5	90.4	82.8	87.8	73.1	62.2	92.6	94.1	60.9	94.5	72.2	77.2	92.7	90.8	76.4
HB3 (5)	4.5	9.6	17.2	12.2	26.9	37.8	7.4	5.9	39.1	5.5	27.8	22.8	7.3	9.2	23.6
3D7 (90)	87.6	88.0	82.3	80.4	66.0	71.7	82.7	86.4	73.3	90.6	76.4	72.5	88.7	83.8	60.7
HB3 (10)	12.4	12.0	17.7	19.6	34.0	28.9	17.3	13.6	26.7	9.4	23.6	27.5	11.3	16.2	39.3
3D7 (80)	77.9	77.7	81.8	73.5	68.9	66.9	74.9	73.7	64.4	84	67.9	67.2	76.4	70.5	59.4
HB3 (20)	22.1	22.3	18.2	26.5	31.1	33.1	25.1	26.3	35.6	16	32.1	32.8	23.6	29.5	40.6

Measured Proportions of SNPs at each marker locus

Calculated values for each clone	pf-01-205304			pf-03-350224			pf-03-842383			pf-10-658467			pf-10-992433		
	0.10	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**
3D7 (20)	46.4	60.0	48.4	51.2	57.8	38.7	59.2	41.7	58.0	47.4	48.2	38.4	48.1	36.0	51.7
HB3 (80)	53.6	40.0	51.6	48.8	42.2	61.3	40.8	58.3	42.0	52.6	51.8	61.6	51.9	64.0	48.3
3D7 (10)	34.6	30.7	37.3	33.2	44.3	31.6	35.7	36.8	37.0	39.6	36.8	35.6	33.9	36.4	50.6
HB3 (90)	65.4	69.3	62.7	66.8	55.7	68.4	64.3	63.2	63.0	60.4	63.2	64.4	66.1	63.6	49.4
3D7 (5)	18.6	27.5	39.1	18.5	41.0	33.2	21.5	34.0	41.6	31.9	27.1	54.5	22.8	20.4	41.4
HB3 (95)	81.4	72.5	60.9	81.5	59.0	66.8	78.5	66.0	58.4	68.1	72.9	45.5	77.2	79.6	58.6
3D7 (2)	12.9	18.1	33.4	17	40.8	39.9	19.7	26.2	38.3	20.6	23.7	30.4	16.2	20.4	38.5
HB3 (98)	87.1	81.9	66.6	83	59.2	60.1	80.3	73.8	61.7	79.4	76.3	69.6	83.8	79.6	61.5
3D7 (0)	7.4	5.2	22.9	7.4	39.4	42	11.2	11.8	31.7	17.3	22.9	28.5	6.8	15.2	32.5
HB3 (100)	92.6	94.8	77.1	92.6	60.6	58	88.8	88.2	68.3	82.7	77.1	71.5	93.2	84.8	67.5

(b)

Measured Proportions of SNPs at each marker locus

Calculated values for each clone	pf-01-393153			pf-01-411796			pf-03-138-7			pf-03-350224		
	0.10	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**
3D7 (100)	100	81.2	70.6	100	80.6	84.0	95.6	83.4	66.2	94.7	88.2	70.1
DD2 (0)	0	18.8	29.4	0	19.4	16.0	4.4	16.6	33.8	5.3	11.8	29.9
3D7 (98)	98.1	79.7	63.6	95.6	76.4	61.2	94.6	80.4	71.7	91.4	80.5	50.4
DD2 (2)	1.9	20.3	36.4	4.4	23.6	38.8	5.4	19.6	28.3	8.6	19.5	49.6
3D7 (95)	95.3	74.6	60.9	95.4	73.8	67.7	90.4	83.4	63.9	89.5	74.7	55
DD2 (5)	4.7	25.4	39.1	4.6	26.2	32.3	9.6	16.6	36.1	10.5	25.3	45
3D7 (90)	93.1	71.5	63.4	88.5	55.8	78.1	87.2	80.1	55.2	86	68.6	76.3
DD2 (10)	6.9	28.5	36.6	11.5	44.2	21.9	12.8	19.9	44.8	14	31.4	23.7
3D7 (80)	83.1	63.9	51.7	79.8	50.3	64.0	84.2	65.5	61.9	72.1	62.9	61.5
DD2 (20)	16.9	36.1	48.3	20.2	49.7	36.0	15.8	34.5	38.1	27.9	37.1	38.5
3D7 (50)	56.5	52.0	44.8	48.9	45.0	42.7	47.7	49.3	55.8	46.9	61.8	53.4
DD2 (50)	43.5	48.0	55.2	51.1	55.0	57.3	52.3	50.7	44.2	53.1	38.2	46.6
3D7 (20)	29.6	38.9	36.0	25.2	38.1	28.6	22.3	43.1	30.5	17.9	58.3	29.4
DD2 (80)	70.4	61.1	64.0	74.8	61.9	71.4	77.7	56.9	69.5	82.1	41.7	70.6
3D7 (10)	19.5	40.2	37.6	14.7	36.8	20.8	13.8	46.4	48.5	7.8	45.7	48.9
DD2 (90)	80.5	59.8	62.4	85.3	63.2	79.2	86.2	53.6	51.5	92.2	54.3	51.1
3D7 (5)	16.2	39.1	39.8	14.2	26.4	37.7	10.7	37.8	43.8	2.2	40.8	37.6
DD2 (95)	83.8	60.9	60.2	86.8	73.6	62.3	89.3	62.2	56.2	97.8	59.2	62.4

Measured Proportions of SNPs at each marker locus

Calculated values for each clone	pf-01-393153			pf-01-411796			pf-03-138-7			pf-03-350224		
	0.10	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**
3D7 (0)	12.5	19.3	40.5	8.3	29.2	5.6	2.5	21.1	40.3	1.8	26.7	20.5
DD2 (100)	87.5	80.7	59.5	91.7	70.8	94.4	97.5	78.9	59.7	98.2	73.3	79.5

(c)

Calculated values for each clone	Measured Proportions of SNPs at each marker locus								
	pf-03-138-7			pf-03-350224			pf-10-658467		
	0.10	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**
3D7 (100)	98.9	90.2	83.4	96.7	93.0	82.2	92.6	90.6	82.3
7G8 (0)	1.1	9.8	16.6	3.3	7.0	17.8	7.4	9.4	17.7
3D7 (98)	96.8	92.0	83.6	93.3	89.4	90.2	92	85.5	82.5
7G8 (2)	3.2	8.0	16.4	6.7	10.6	9.8	8	14.5	17.5
3D7 (95)	96.9	92.2	90.7	94	76.5	85.6	89.3	82.8	76.9
7G8 (5)	3.1	7.8	9.3	6	23.5	14.4	10.7	17.2	23.1
3D7 (90)	93.6	89.7	80.1	93.8	72.3	78.5	91.3	78.4	83.8
7G8 (10)	6.4	10.3	19.9	6.2	27.7	21.5	8.7	21.6	16.2
3D7 (80)	91.4	72.2	65.5	88.5	68.6	68.8	89.1	77.8	77.4
7G8 (20)	8.6	27.8	34.5	11.5	31.4	31.2	10.9	22.2	22.6
3D7 (50)	70.3	65.5	49.3	69.9	65.3	69	60.3	68.1	70.6
7G8 (50)	29.7	34.5	50.7	30.1	34.7	31	39.7	31.9	29.4
3D7 (20)	46.6	51.9	36.4	40.5	58.6	44.4	49.3	42.5	32.7
7G8 (80)	53.4	48.1	63.6	59.5	41.4	55.6	50.7	57.5	67.3
3D7 (10)	44.1	31.9	20.1	36.4	55.7	46	19.9	33.1	42.4
7G8 (90)	55.9	68.1	79.9	63.6	44.3	54	80.1	66.9	57.6
3D7 (5)	23.2	46.0	19.8	22.4	48.1	22.4	9.8	17.7	17.7
7G8 (95)	76.8	54.0	80.2	77.6	51.9	77.6	90.2	82.3	82.3

Measured Proportions of SNPs at each marker locus

Calculated values for each clone	pf-03-138-7			pf-03-350224			pf-10-658467		
	0.10	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**	**0.10**	**0.010**	**0.001**
3D7 (0)	18.4	18.8	48.5	8.7	32.3	20.4	0.4	7.0	25.7
7G8 (100)	81.6	81.2	51.5	91.3	67.7	79.6	99.6	93.0	74.3

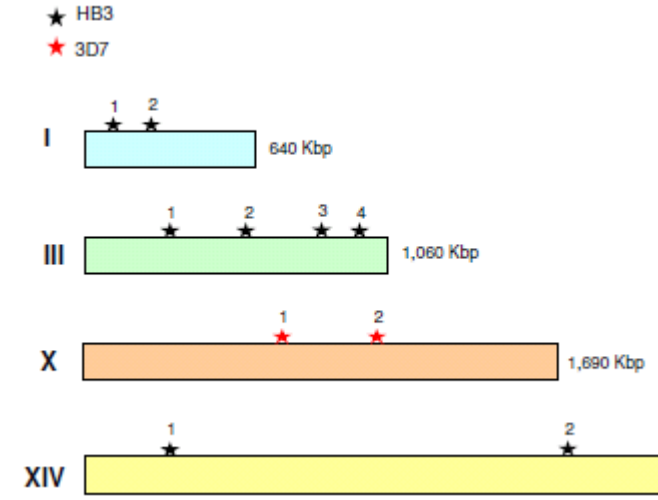
The results in **Table 4.3 a, b & c** indicate that it is possible in most cases to identify the major clone within the limits previously described in **Section 4.2**. This was demonstrated clearly at 0.1% parasitaemias, however, at 0.01% and 0.001% parasitaemias, only the major clone could be identified in most cases (see **Table 4.3b** & **Table 4.3c**). Thus parasite genotypes can more often than not be clearly deduced from the major clone

As already observed in **Table 4.1 a, b & c**, Pyrosequencing™ has limits in detecting clones occurring at low proportions (minor clones) accurately. The ability of the technique to accurately assign proportions to the minor clones was further compromised by reduced parasitaemias. The results in **Table 4.3 a, b & c** show that the minor clones could be identified in most cases within the limits of detection previously described at parasitaemias of 0.1% but at 0.01% and 0.001% parasitaemia tested in this stud but, it was almost impossible to identify the minor clone. From this results it can be concluded that Pyrosequencing™, is in most cases able to accurately assign proportions to samples having parasitaemias of approximately 0.1% or higher and, parasite genotypes present in these mixtures can be obtained, but within the limits of detection already described in **Section 4.2.**

4.5. Identification of the parasite genotypes derived from the semi-cloned 3D7 and HB3 cross progeny

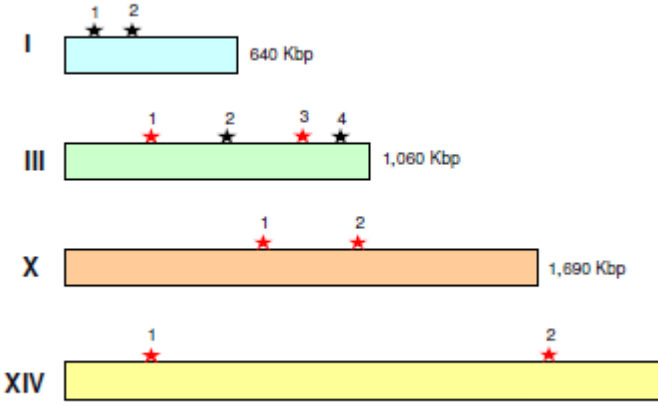
The 3D7 and HB3 cross progeny, semi-cloned as described in **Chapter 2**, were analysed by Pyrosequencing™ using the marker assays identified in **Chapter 3** and known to distinguish the genetically distinct clones 3D7 and HB3. The results are shown below in **Figure 4.3**.

The results in **Figure 4.3** comprise 10 marker assays that have already been identified in **Table 3.2**, **Chapter 3** to contain SNPs that distinguish the genetically distinct clones 3D7 and HB3.



Sample 1

	3D7		HB3		Genotype
Marker	SNP	Quantitation	SNP	Quantitation	
Pf-01-101502	G	0	T	100	T
Pf-01-205304	G	0	A	100	A
Pf-03-350224	C	0	T	100	T
Pf-03-613146	G	0	A	100	A
Pf-03-842383	A	3.4	C	96.6	C
Pf-03-967838	C	0	A	100	A
Pf-10-658467	C	100	T	0	C
Pf-10-992433	A	100	G	0	A
Pf-14-279668	A	0	T	100	T
Pf-14-1710196	A	3.7	T	96.3	T



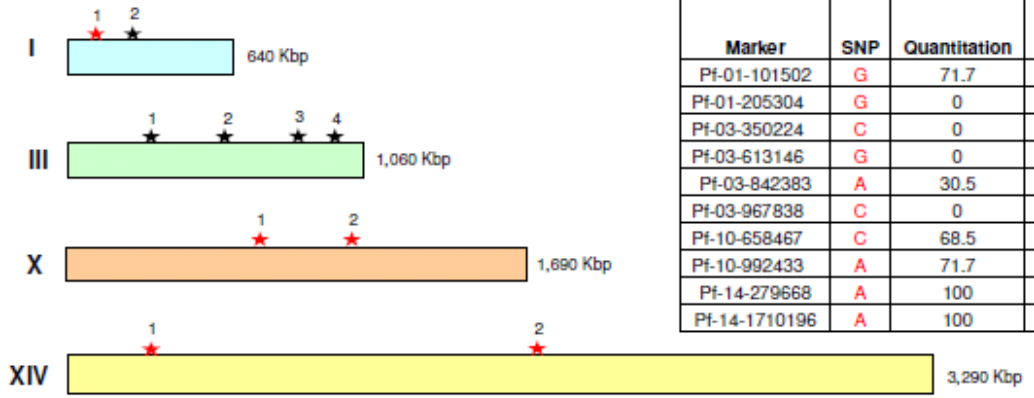
Sample 2

	3D7		HB3		
Marker	SNP	Quantitation	SNP	Quantitation	Genotype
Pf-01-101502	G	0	T	100	T
Pf-01-205304	G	0	A	100	A
Pf-03-350224	C	98.1	T	1.9	C
Pf-03-613146	G	0	A	100	A
Pf-03-842383	A	88.2	C	11.8	A
Pf-03-967838	C	0	A	100	A
Pf-10-658467	C	100	T	0	C
Pf-10-992433	A	100	G	0	A
Pf-14-279668	A	100	T	0	A
Pf-14-1710196	A	100	T	0	A

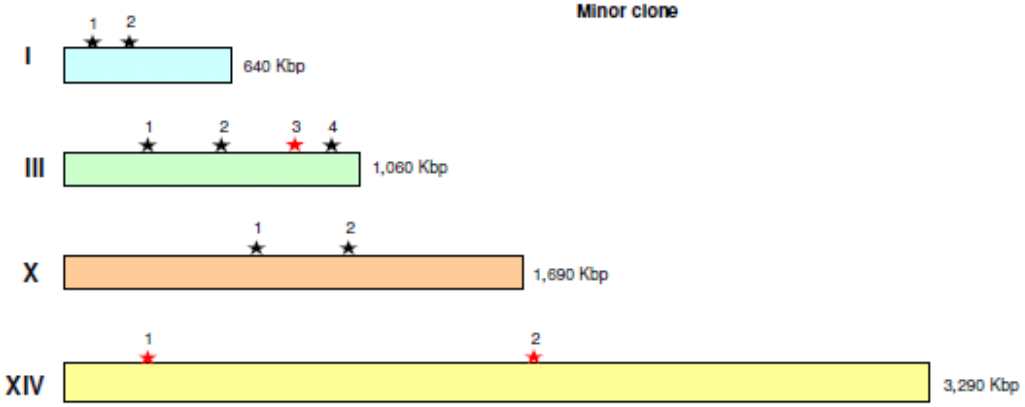
Major clone

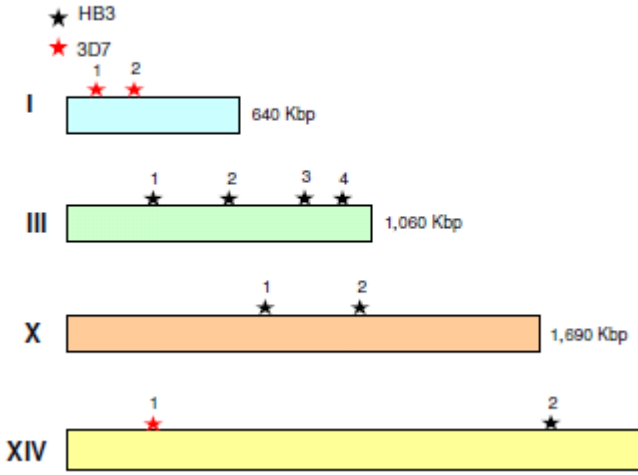
Sample 3

Marker	SNP	3D7 Quantitation	SNP	HB3 Quantitation	Genotype Clone 1 (Major)	Genotype Clone 2 (Minor)
Pf-01-101502	G	71.7	T	28.3	G	T
Pf-01-205304	G	0	A	100	A	A
Pf-03-350224	C	0	T	100	T	T
Pf-03-613146	G	0	A	100	A	A
Pf-03-842383	A	30.5	C	69.5	C	A
Pf-03-967838	C	0	A	100	A	A
Pf-10-658467	C	68.5	T	31.5	C	T
Pf-10-992433	A	71.7	G	28.3	A	G
Pf-14-279668	A	100	T	0	A	A
Pf-14-1710196	A	100	T	0	A	A



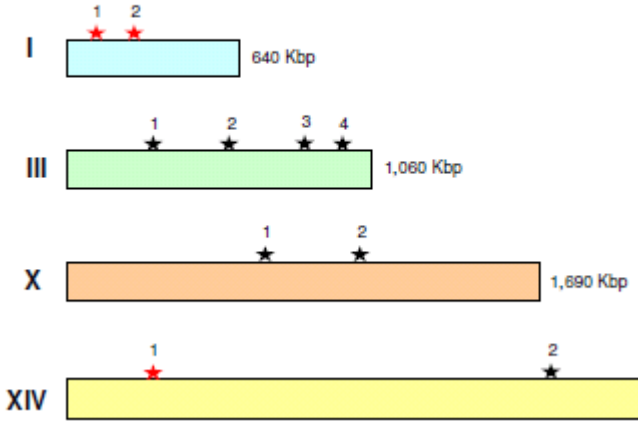
Minor clone





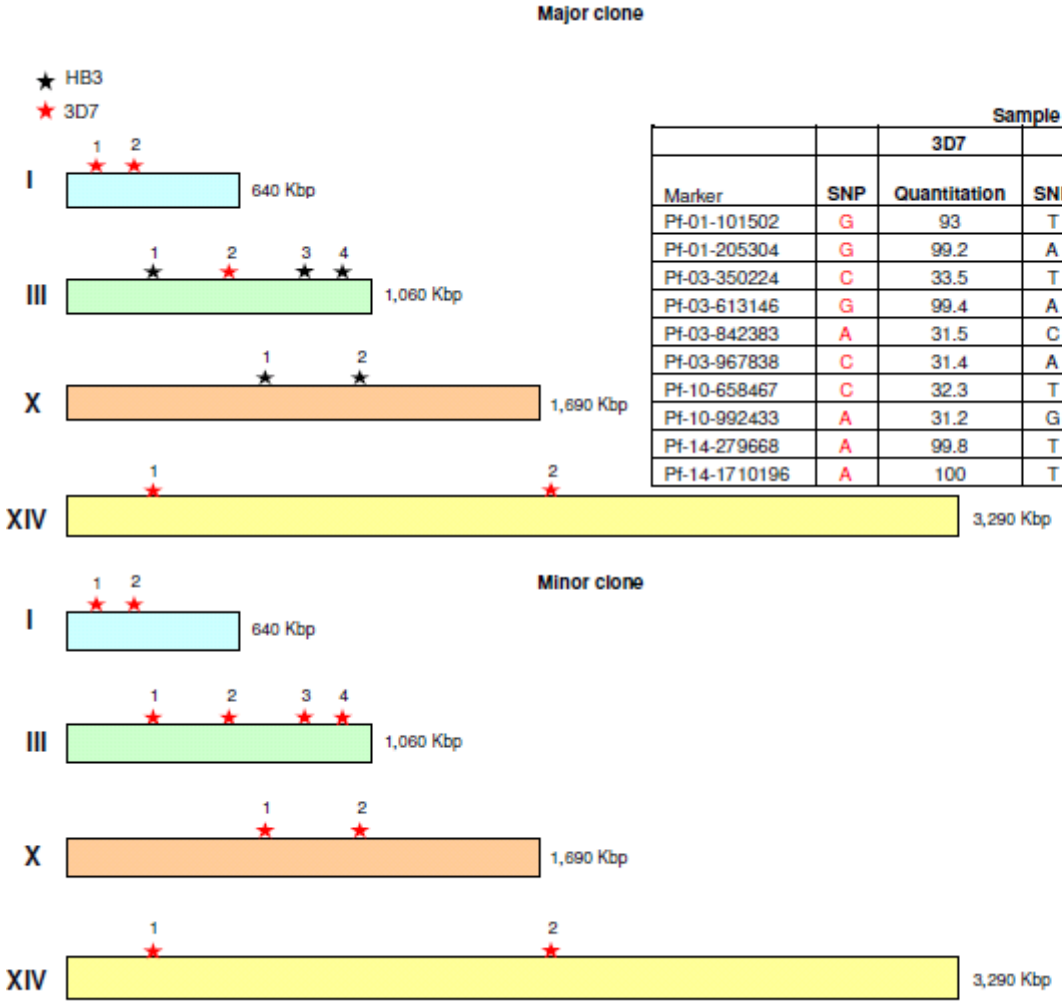
Sample 4

		3D7		HB3	
Marker	SNP	Quantitation	SNP	Quantitation	Genotype
Pf-01-101502	G	92.3	T	7.7	G
Pf-01-205304	G	99.2	A	0.8	G
Pf-03-350224	C	0	T	100	T
Pf-03-613146	G	0	A	100	A
Pf-03-842383	A	3.1	C	96.9	C
Pf-03-967838	C	0	A	100	A
Pf-10-658467	C	0	T	100	T
Pf-10-992433	A	0.8	G	99.2	G
Pf-14-279668	A	96.5	T	3.5	A
Pf-14-1710196	A	3.2	T	96.8	T



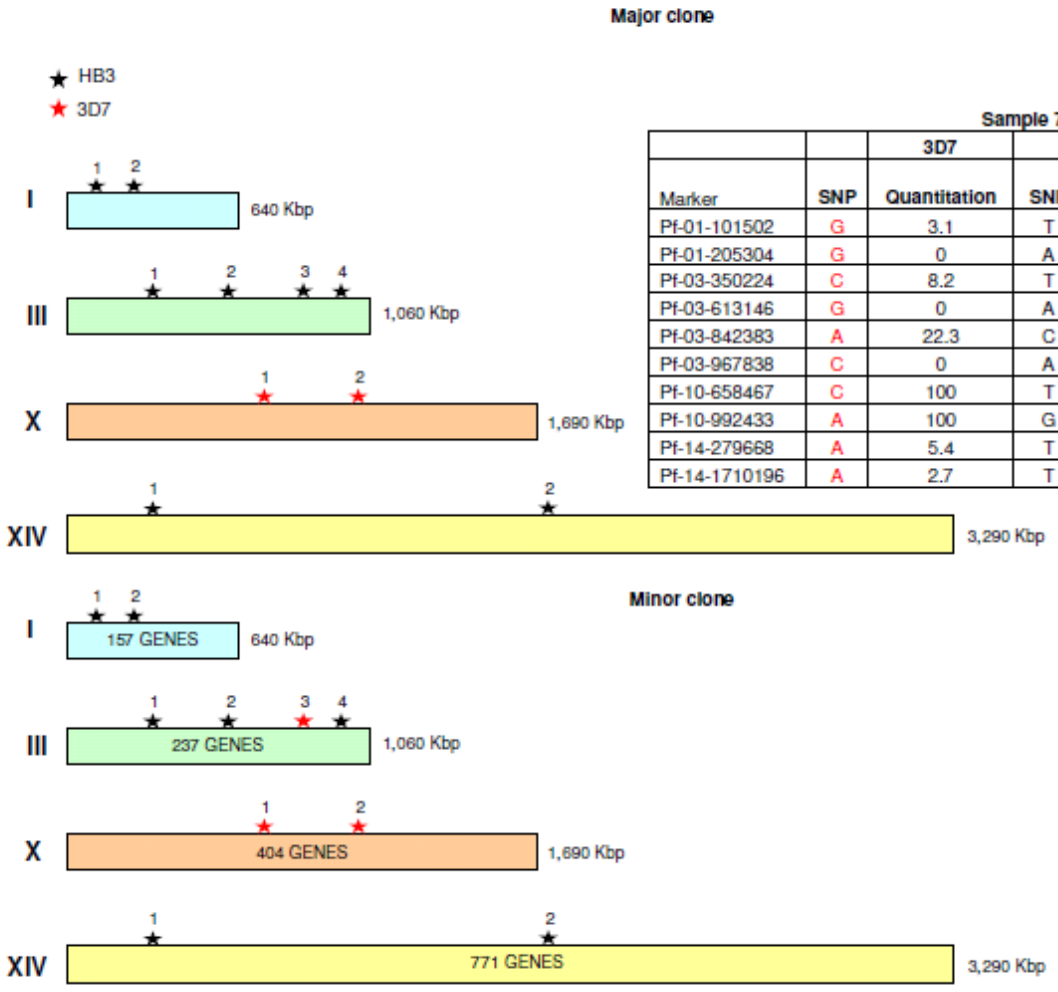
Sample 5

		3D7		HB3	
Marker	SNP	Quantitation	SNP	Quantitation	Genotype
Pf-01-101502	G	94.7	T	5.3	G
Pf-01-205304	G	100	A	0	G
Pf-03-350224	C	0	T	100	T
Pf-03-613146	G	0	A	100	A
Pf-03-842383	A	1.8	C	98.2	C
Pf-03-967838	C	0	A	100	A
Pf-10-658467	C	3	T	97	T
Pf-10-992433	A	4.8	G	95.2	G
Pf-14-279668	A	98.9	T	1.1	A
Pf-14-1710196	A	2.8	T	97.2	T



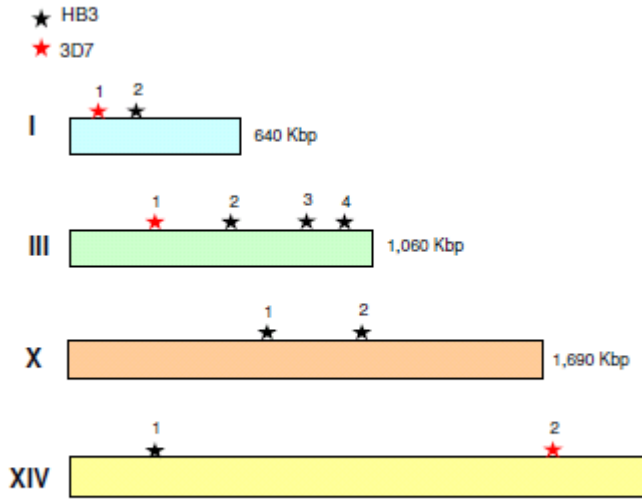
Sample 6

		3D7		HB3	Genotype	
Marker	SNP	Quantitation	SNP	Quantitation	Clone 1 (Major)	Clone 2 (Minor)
Pf-01-101502	G	93	T	7	G	G
Pf-01-205304	G	99.2	A	0.8	G	G
Pf-03-350224	C	33.5	T	66.5	T	C
Pf-03-613146	G	99.4	A	0.6	G	G
Pf-03-842383	A	31.5	C	68.5	C	A
Pf-03-967838	C	31.4	A	68.6	A	C
Pf-10-658467	C	32.3	T	67.7	T	C
Pf-10-992433	A	31.2	G	68.8	G	A
Pf-14-279668	A	99.8	T	0.2	A	A
Pf-14-1710196	A	100	T	0	A	A



Sample 7

		3D7		HB3	Genotype	
Marker	SNP	Quantitation	SNP	Quantitation	Clone 1 (Major)	Clone 2 (Minor)
Pf-01-101502	G	3.1	T	96.9	T	T
Pf-01-205304	G	0	A	100	A	A
Pf-03-350224	C	8.2	T	91.8	T	T
Pf-03-613146	G	0	A	100	A	A
Pf-03-842383	A	22.3	C	77.7	C	A
Pf-03-967838	C	0	A	100	A	A
Pf-10-658467	C	100	T	0	C	C
Pf-10-992433	A	100	G	0	A	A
Pf-14-279668	A	5.4	T	94.6	T	T
Pf-14-1710196	A	2.7	T	97.3	T	T



Sample 8

		3D7		HB3	
Marker	SNP	Quantitation	SNP	Quantitation	Genotype
Pf-01-101502	G	93.7	T	6.3	G
Pf-01-205304	G	0	A	100	A
Pf-03-350224	C	99.1	T	0.9	C
Pf-03-613146	G	0	A	100	A
Pf-03-842383	A	7.5	C	92.5	C
Pf-03-967838	C	0	A	100	A
Pf-10-658467	C	1.5	T	98.5	T
Pf-10-992433	A	2.1	G	97.9	G
Pf-14-279668	A	0.6	T	99.4	T
Pf-14-1710196	A	100	T	0	A

Figure 4.3 Schematic representation of the 8 samples of the 3D7 and HB3 semi-cloned cross progeny analysed by Pyrosequencing™. The coloured bars represent chromosomes I, III, X and XIV with their sizes given in kilo base pairs (Kbp) on the far right. The stars with the numbers 1, 2…etc. above them represent the SNPs (marker loci) on the approximate region they occupy in the chromosomes. These are given as follows:

- Chromosome I – 1. Pf-01-101502, 2. Pf-01-205304
- Chromosome III – 1. Pf-03-305204, 2. Pf-03-613146, 3. Pf-03-842383, 4. Pf-03-967838
- Chromosome X – 1. Pf-10-658467, 2. Pf-10-992433
- Chromosome XIV- 1. Pf-14-279668, 2. Pf-14-1710196

The tables on the right show the samples analysed (Samples 1-8), the marker loci in each chromosome, the quantitation values assigned by Pyrosequencing™ at each SNP (marker locus) and the resulting genotypes. The 3D7 parental clone genotypes are given in red and the HB3 parental clone genotypes are given in black. The resulting genotypes are visually represented in the figures. The locus defined by the parental clone 3D7 is represented by a red star and the locus defined by the parental clone HB3 by a black star.

Following analysis by Pyrosequencing™, the SNPs identified at each marker locus for 3D7 (red) and HB3 (black) and their proportions are given in the tables within **Figure 4.3**. The parasite genotypes of each sample analysed are also shown in the tables and in cases where two clones are represented each is defined as the clone occurring at a high proportion (the major clone or clone 1) and the one occurring at a lower proportion (minor clone or clone 2). The marker loci are mapped at the approximate positions they occupy in the chromosomes as shown in the figures on the left of **Figure 4.3**. The SNP identified by Pyrosequencing™ representing either of the two clones (3D7 or HB3) at each marker locus in the figures is represented by 3D7 (red star) or HB3 (black star).

The results in **Figure 4.3** show that following semi-cloning (described in **Figure 2.1**), either one or two genetically distinct clones were recovered from the 8 wells (Samples 1-8). Samples 1, 2, 4, 5 and 8 each contained one parasite genotype and samples 3, 6 and 7 contained two parasite genotypes. The criterion for identifying the genotypes is as described in **Chapter 4**. For the samples having one parasite genotype identified, the low proportions (indictive of a minor clone) assigned by Pyrosequencing™ machine were ignored as they did not conform to the limits set in the identification of minor clones (**Section 4.2.1**). This is based on the premise that, parasite genotypes are derived from the estimated SNP proportions and that, in a two genotype infection, SNPs occurring at a higher proportion constitute one parasite genotype, while those occurring at a lower proportion constitute the other. This was demonstrated in samples 3, 6 and 7 which revealed major clones represented at ~70% and minor clones at ~30%. Interestingly, Samples 4 & 5 and Samples 1 & 7 (Clone 1) consisted of indistinguishable genotypes according to the markers used.

It is also noteworthy that all of the genotypes identified from the cross progeny (**Figure 4.3)** were recombinants except Sample 6 that had one parental genotype (3D7) as a minor clone. These results are analogous to what was observed by Walliker *et al*., 1987 where, a high proportion of recombinants was also observed in the original analysis of this cross progeny.

4.6. Estimating the genetic relatedness of the 3D7 and HB3 cross progeny

From the formula in **Section 4.1.3**, a GR value of 1 represents "total genetic relatedness", while 0 represents "no genetic relatedness". It is noteworthy that most of the population genetic software use the parameter GD and not GR for analyses. Either GD or GR can be used to define levels of genetic diversity in *P. falciparum*. The results of the typical genetic relatedness matrix (or genetic distance obtained by subtracting the values in the given matrix with 1) obtained from pairwise comparisons of genotypes in the estimation of genetic diversity for subsequent population genetics analyses are summarised in **Table 4.3**.

Table 4.3 A matrix containing a summary of pair-wise allele comparisons across all loci analysed for the 3D7 and HB3 cross progeny. All the parental clones and the cross progeny are given below. Where there is either (a) or (b) signifies the major and minor clones for example 3a and 3b. The average GRs between the parental and cross progeny genotypes and, among the cross progeny genotypes are also given.

Samples	3D7	HB3	1	2	3a	3b	4	5	6a	6b	7a	7b	8
1	0.2	0.8											
2	0.6	0.4	0.6										
3a	0.5	0.5	0.7	0.7									
3b	0.3	0.7	0.5	0.7	0.6								
4	0.3	0.7	0.5	0.3	0.6	0.6							
5	0.3	0.7	0.5	0.3	0.6	0.6	1						
6a	0.5	0.5	0.3	0.7	0.6	0.6	0.8	0.8					
6b	1	0	0.2	0.6	0.5	0.3	0.3	0.3	0.5				
7a	0.2	0.8	1	0.6	0.7	0.5	0.5	0.5	0.3	0.2			
7b	0.3	0.7	0.9	0.7	0.6	0.6	0.4	0.4	0.2	0.3	0.9		
8	0.2	0.8	0.5	0.5	0.6	0.6	0.6	0.6	0.6	0.3	0.5	0.4	
Average GR	**0.4**	**0.6**	**0.53**										

Assuming random fertilization of gametes and no bias in survival, the resulting first generation progeny would be expected to comprise 50% of parental types and 50% of recombinants. The expected average GR would therefore be 0.5. The results in **Table 4.3** show that the average GR between the cross progeny genotypes and each of the parental genotypes was found to be 0.4 and 0.6 for 3D7 and HB3 respectively

and, the average GR among the cross progeny was 0.53. These observed GR values are close to the expected GR of 0.5 in conformity to the null hypothesis.

4.7. Discussion

The aim of this chapter was to test the reliability of Pyrosequencing™ in accurately detecting and quantifying SNPs representing genetically distinct clones from laboratory prepared parasite mixtures and subsequently identifying the parasite genotypes represented therein. It has been shown that the technique is indeed able to accurately assign proportions to the SNPs and parasite genotypes could be identified from the parasite mixtures on the premise that SNPs occurring at similar proportions belong to the same genotype.

In **Table 4.1** results, the median values were used as opposed to the mean of the overall measurements, to accommodate the outlier values observed in the triplicate measurements. The lack of replicability in the observed proportions was due to pipetting errors when preparing the Pyrosequencing™ assays. This method is prone to pipetting errors as with most DNA analysis methods so care should be taken when carrying out the analysis.

It is also noteworthy that inconsistencies were observed in the values assigned by the Pyrosequencing™ vis-à-vis the calculated proportions of the laboratory prepared clone mixtures (**Figure 4.1 a, b & c)**. It is plausible that the discrepancy arose in the preparation of the tested parasite clones. To ascertain this, it is recommended that fresh parasite clone mixtures with the same proportions should be prepared and the experiment repeated.

The experiments in this Chapter were conducted to validate Pyrosequencing™ for use in analysis of P. *falciparum* field isolates which have unknown parasite genotypes and at varying proportions. The technique has been shown to have a limit of detection for clones occurring in low proportions (minor clones) but in most cases it should be possible to identify at least one parasite genotype (derived from the major clone) in mixed clone infections. These results are in concordance with another study investigating the different types of *Toxoplasma gondii* wild isolates by Pyrosequencing™. The analysis revealed that at least two genotypes could be identified and discriminated in a mixed genotype infection but the technique was limited in identifying more than two genotypes (Sreekumar *et al.*, 2005; Edvinsson *et al*., 2007).

This is a major setback in using this technique especially when analysing African *P. falciparum* isolates where as many as 8 clones have been found in individual infections (**Table 1.1, Chapter 1**). This is however an overall problem when using molecular genotyping technology in the detection of clonal genotypes occurring in wild isolates. To circumvent it, Nkhoma *et al*., (2012) for example cultured parasites from field collections and cloned them by limiting dilution. This has two main purposes; first of all it is a DNA enrichment process for sequencing purposes. Secondly it increases the chances of clone detection by ensuring that only one parasite per well is recovered for genotyping. It is highly effective and in this study a maximum of 6 genetically distinct clones per infection were detected in Malawi but using the Illumina genotyping platform.

The detection of the major clone in most cases might be due to the PCR technique selectively amplifying the most abundant clones leading to the minority clones being

missed out. This is a common problem in PCR based genotyping technologies. However, the parasite biology of *P. falciparum* further exacerbates the situation as in most cases the major clones are fast growers and out-compete the minor clones rendering them undetectable. This has been observed in the rodent malaria parasite *P. chabaudi adami* where the DS strain is fast growing while the DK strain is slow growing (Gadsby & Carter, 2008).

This Chapter also outlined that Pyrosequencing™ not only has limited accuracy when assigning proportions in clones occurring at low proportions but also when dealing with isolates present at parasitaemias less than 0.01%. *P. falciparum* field isolates usually occur at parasitaemias greater than 0.1% in most of the malaria positive symptomatic cases. It is also possible to encounter malaria positive asymptomatic cases at less than 0.01% parasitaemias. Thus, when analysing *P. falciparum* field isolates by Pyrosequencing™, the technique's limit of detection is parasitaemias >0.1% as it is possible in most cases to identify parasite genotypes from samples occurring at these parasitaemias. This is a major setback as it means that when dealing with asymptomatic cases, many parasite clones might remain undetected.

Attempts to genotype *P. falciparum* wild isolates can be further hampered by the existence of these parasites at low parasitaemia levels in peripheral blood. This is because they sequester reducing them to undetectable levels. Sequestration has been observed to occur during the 48hours cyclic asexual cycle in the peripheral blood and lasts for approximately 24 – 28 hours (Koepfli *et al*., 2011). To enhance the probability of developing accurate genotypic profiles in multiclonal infections, repeated sampling every 24 hours coinciding with emergence of sequestered

parasites to the blood has been recommended (Koepfli *et al*., 2011). This sampling strategy is however impractical when working with human subjects.

In this chapter, a genetic cross between the known genetically characterised clones 3D7 and HB3 was analysed. The results showed that both parental genotypes were represented in the cross progeny and interestingly one parental genotype was also identified in the cross progeny (**Figure 4.3**). The level of genetic relatedness between the parental genotypes and the cross progeny genotypes as well as among the cross progeny genotypes was investigated and was found to be close to 0.5 as expected (**Table 4.3**).

The strategy outlined in this section, whereby the genetic diversity of the malaria parasites is summarised in a genetic distance, will be utilised in later chapters of this thesis for the subsequent analysis using genetic distance based analysis.

4.7.1. Conclusion

In this chapter the identified SNPs in **Chapter 3** have been used in the identification and quantification of parasite clones represented in laboratory prepared mixtures of parasite clones at known proportions. In addition, the SNPs have also been tested using progeny from a cross between the genetically distinct clones 3D7 & HB3 with each of the clones present at unknown proportions. These tests have been performed to validate use of these SNPs in analysis of *P. falciparum* field isolates. They have facilitated the understanding of the technique's limits of detection which will be useful in the interpretation of genotypic information obtained from the wild isolates.

5. THE ANALYSIS OF *P. FALCIPARUM* FIELD ISOLATES

5.1. Introduction

The existence of *P. falciparum* as complex mixtures of genetically distinct clones requires the use of genetic markers to identify and distinguish the clones present within human host infections. The choice of genetic markers to use depends on the study of interest. For example, to generate *P. falciparum* drug resistance profiles, polymorphic genetic markers under selection pressure imposed by drugs located at genes of interest are used (reviewed by Wongsrichanalai *et al*., 2002). On the other hand, in studies of population genetic structure, neutral genetic markers are recommended to ensure that the observed population structure is due to gene flow patterns and not by natural selection or immune processes (Anderson *et al*., 1999; Arez & Rosário, 2008).

This study entailed an attempt to determine population divergence and gene flow patterns of malaria parasites within small geographic settings, the genetic markers of choice were therefore SNPs and, microsatellites were included in the analysis as additional markers for the identification of mixed clone infections.

5.1.1. SNPs

The use of SNPs as genetic markers has garnered tremendous interest (Morin *et al.*, 2004) driven mostly by the human genome project (see

http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml). The success of this project in identification of geographically distinct human populations using SNPs has paved the way for advances in sequencing technology (França, *et al*., 2002; Metzker, 2005; Metzker, 2010; Zhou *et al*., 2010; Shen *et al*., 2010). The existence of these technologies has led to the extensive use of SNPs across different species (Kumar *et al*., 2012).

SNPs are numerous across genomes e.g. genome-wide comparisons of cattle (The bovine hapmap consortium, 2009) or in chicken (Groenen *et al*., 2011) breeds rendering them highly informative. In addition, genome wide SNP analysis enables the detection of low genetic diversity occurring within and between organisms. For example, in *P. falciparum* studies, SNPs are of significant use in identifying genetically distinct clones present in mixed infections (Anderson *et al*., 1999; Campino *et al.*, 2011; Juliano *et al*., 2010). In high malaria transmission areas where extensive genetic recombination takes place, parasites are highly diverse and a handful of genetic markers are sufficient to decipher the genetically distinct clones. However, in regions of low malaria transmission where parasite infections are sporadic and widespread such that parasites are almost genetically identical discrimination among clonal genotypes is difficult (Bendixen *et al*., 2001) without numerous genetic markers.

5.1.2. Microsatellites

Microsatellites comprise highly variable repetitive motifs occurring predominantly in nuclear genomes of organisms (Selkoe & Toonen, 2006) in both coding and

noncoding regions (Wilder & Hollocher, 2001). Microsatellites initially gained interest due to their role in the occurrence of some human neurological disorders such as fragile X syndrome (reviewed by O'Donnell & Warren, 2002). This paved way for their use in the generation of genetic maps in other organisms such as in Zebra fish (Knapik *et al*., 1998), Soybeans (Cregan *et al*., 1999) and even in *P. falciparum* (Su & Wellems, 1996). They have also been used in genetic parentage analysis and population genetic studies (Oliveira *et al*., 2006).

In population genetic studies, microsatellites are favoured as they are abundant across the genome enabling the analysis of multiple loci (Li *et al*., 2002). This is essential especially in the detection of population structure as different loci exhibit different levels of genetic diversity. Different loci experience random fluctuations in allele frequencies as a result of genetic drift and variation in coalescent history, necessitating the need for multilocus analysis in the determination of population structure between and among subpopulations (Nei, 1973). Microsatellites are also highly polymorphic due to their high rate of mutation (Jarne & Lagoda, 1996; Ellegren, 2004). In addition they have been observed to be selectively neutral (Goldstein & Schlötterer, 1999; Schlötterer, 2000) which is important when studying demographic patterns in populations as selection may confound the interpretation of gene flow patterns between them.

12 microsatellites assumed to be neutral were identified and validated for use in the study of *P. falciparum* populations (Anderson *et al*., 1999). In addition, numerous microsatellites have been identified across the *P. falciparum* genome (Mu *et al*., 2007) and are available in the website: (http://www.plasmodb.org/plasmo/home.jsp).

In this study a total of 8 microsatellites obtained from different chromosomes across the *P. falciparum* genome were analysed. 5 were adapted from Anderson *et al*., (1999) namely: Poly α, TA81, TA40, PFPK2, TA60 and, 3 from (Tsumori *et al*., 2011) i.e. TA17, TA53, TA43. These were used as additional genetic markers to the available SNPs described in previous chapters. The laboratory analysis was carried out by collaborators at Nagasaki University, Japan.

5.1.3. Synopsis of chapter

In this chapter, I determine the informativeness of SNPs and microsatellites in terms of levels of between-clone genetic diversity detected and quantify the genetic variation observed within malaria parasite populations for both genetic markers. For microsatellites, I test the assumption of selective neutrality to identify whether the observed divergence is due to neutral mutations or selection. This is important the observed alleles may have evolved under different forces, ranging from natural selection to genetic drift. Since the purpose of this study is to determine how closely related are parasites transmitted within small spatial settings, the expected variation or lack thereof is highly dependent on gene flow patterns. Any influence of evolutionary forces such as natural selection and gene flow should be deduced prior to interpreting observed patterns of genetic diversity (Jensen *et al*., 2005; Ramírez-Soriano *et al*., 2008).

The variation within and between populations (at a country or village level) will be determined. This is a measure of the level of population structure observed and reported as the Wright's fixation index (F_{ST}). To achieve the objective of this study,

malaria parasites will be partitioned geographically with distances generated randomly. Comparison of these randomly generated subpopulations giving non-significant F_{ST} values will imply that there is extensive gene flow between individual malaria parasites. Geographic partitions with significant F_{ST} values will be indicative of genetic differentiation between the subpopulations.

5.1.4. Aims of this chapter

1. To determine if the genetic markers used are under any form of selection. This is performed in the form of a neutrality test or by detection of linkage disequilibrium (LD) (statistical association of alleles). Loci that are found to be in LD comprise of allelic combinations that occur at frequencies greater than expected under random mating.
2. To determine the degree of genetic markers informativeness following analysis of the field isolates. This will be done by computation of allele frequencies and gene diversity (heterozygosity).
3. Test for significant levels of population genetic structure in parasites to reveal how closely related genetically are the malaria parasites within small geographic settings (geographic distances of <1km).

5.1.5. Null Hypotheses

1. Parasite populations conform to random mating patterns and are not under any form of selection

2. The genetic markers are in linkage equilibrium i.e. constituting alleles that are assorted independently and are therefore not statistically associated.
3. The parasite populations are isolated with no migration between them, are not partitioned into small subpopulations or genetically structured in space.

5.2. Materials and methods

5.2.1.1. Identification of clonal genotypes

The SNPs validated in **Chapter 4** were used to analyse field isolates. Only two clones could be clearly identified in isolates having more than one clone. This is because the proportions assigned by the Pyrosequencer were too close in value to enable the discrimination of 3 or more clones. Further to this, although in **Chapter 4** clones could be identified from laboratory prepared clone mixtures when the major clones proportions were >60% and the minor clones <40%, when dealing with field isolates, the major clone was confirmed only if it occurred at proportions >70% and a minor clone at proportions <30% per isolate. This is because the cut-off point in **Chapter 4** was observed under controlled experimental conditions where only two clones were present per sample analysed whereas in field isolates it is possible to encounter more than two clones. Because of this high stringency in the identification of parasite clones, more sample losses were experienced. In some other cases, two clones observed to occur at proportions close to 50% were also discarded as it was difficult to assemble individual clonal genotypes.

Both SNPs and microsatellites were used to identify clonal genotypes. SNPs and microsatellites are usually analysed separately because they evolve and diversify

under different mutation models. SNPs are analysed using the infinite alleles model (IAM) (Kimura & Crow, 1964), which makes the assumption that each allele generated in a population by mutation arises independently of a previous one that may have occurred at the same site. Microsatellites on the other hand, are analysed using the stepwise mutation model (SMM). Since microsatellites constitute alleles of different sizes due to their tandem repetitive nature, the SMM assumes that each allele generated by mutation is dependent on the state of the previous allele (Kimura & Ohta, 1978). It follows that alleles showing the least difference in size have close genetic relatedness while those that are highly variable between clones are more distantly related.

However, in this study, the mutation rate can be ignored as it involves the determination of parasite genetic relatedness within short time spans such that genetic divergence due to mutations between parasite generations is negligible. Nonetheless, Balloux & Lugon-Moulin, (2002) in their review article suggest that microsatellites have a high mutation rate and the underlying mutation mechanism should not be ignored especially in the determination of genetic differentiation of populations.

5.2.1.2. Test of informativeness

Using the software FSTAT vs 2.9.3 (Goudet, 2001), the SNPs and microsatellites were tested for their level of informativeness by scoring allele frequencies per locus. The number of alleles and the allelic richness were also reported. Allelic richness is a measure of genetic diversity, constitutes the mean number of alleles per sample and is dependent on the sample sizes as large samples are expected to have more alleles

than small samples. To circumvent the problem of unequal sampling, a rarefaction method is used to allow the comparison of genetic diversity across different samples sizes. Rarefaction involves randomly resampling alleles from the whole data set and deriving the expected number of alleles from a given small sample set (in this case fixed for the country with the smallest number of genotypes) (see review by Leberg, 2002). In FSTAT, from a diploid set (*P. falciparum* is haploid, diploid genotypes were obtained by coding the same allele twice) the sample size is fixed for the population with the smallest number of individuals typed (in this case Cameroon with 12 genotypes).

5.2.1.3. Genetic diversity

The overall genetic diversity (heterozygosity) in each of the geographic locations was computed. FSTAT calculates gene diversity or the expected heterozygosity (Hexp) (under HW random mating expectations as the null hypothesis, H_0). See formula in **Appendix G**.

F statistics were developed to determine reductions of heterozygosity within subpopulations due to non-random mating and among subpopulations due to structuring. In haploid populations, $\mathbf{F_{ST}}$ is the mean reduction of expected heterozygosity among subpopulations relative to the total expected heterozygosity of the population due to structuring [**Hexp (t) – Hexp (s) / Hexp (t)**].

Given that different loci are under different forces (e.g. natural selection or genetic drift) other than gene flow effects, multiple loci are used to estimate average F

statistics values as this reduces the bias occurring on individual loci due to other forces acting on them. Thus, the value H_{sk} or Hexp (see **Appendix G**) is an unbiased estimate across different loci and across different individuals and is then used in the determination of F statistics values.

FSTAT computes an estimate of F_{ST} based on multilocus averages of the expected heterozygosities known as G_{ST} or Nei's F statistics (Nei, 1973). The formulae for this statistic are given in **Appendix G**.

FSTAT also computes Weir and Cockerham (1984) F statistics also known as θ. Unlike the Nei's estimate which weights all samples equally irrespective of the sample size differences, θ weights allele frequencies according to the sizes of the analysed samples. In addition, the test provides standard error estimates and confidence intervals by the randomisation and resampling method. This test is therefore considered to be more robust than Nei's estimate.

FSTAT tests for significance of F_{ST} values by the randomisation of genotypes between pairs of samples. 1000 permutations were performed and the log likelihood G-statistic (Goudet, *et al*., 1996; Petit *et al*., 2001) used as an overall test statistic to validate the P values. The tests performed in FSTAT are usually designed for one locus therefore the G-statistic is used to obtain a P value based on the overall estimate of the combined loci in the data set. The G-statistic performs a randomisation test on the genotypes among subpopulations and not alleles. This statistic is considered powerful as there is a possibility that individual alleles are correlated within individuals hence the use of genotypes for randomisation (Goudet *et al*., 1996).

Since microsatellites are believed to undergo a step-wise mutation model (Slatkin, 1995), the genetic differentiation parameter under this model R_{ST}, was estimated. This parameter takes into consideration the variation of microsatellites' allele sizes. FSTAT estimates R_{ST} for each locus following Rousset (1996). This estimator does not depend on the number of samples. It also outputs the different components of variance of allele size: Va for among samples, Vb for among individuals within samples and Vw for within individuals. Vt = Va + Vb + Vw is thus an unbiased estimate of the total overall variance in allele size. The R_{ST} estimates recommended by Goodman (1997) whereby the variance of allele sizes is standardised by using the centred normalised allele size are reported here.

5.2.1.4. Linkage disequilibrium

Microsatellites only were analysed for linkage disequilibrium. Linkage disequilibrium is the non-random association of alleles across loci and can occur as a result of a range of processes including population sub-structure and selection. To test for LD, the clonal genotypes that were identical at all loci were first omitted from the data set and then included in a second analysis and the results compared. The software LIAN vs 3.5 (Haubold & Hudson, 2000) available as a web interface, was used for the LD analysis. LIAN detects this by first determining the number of loci at which the genotypes or haplotypes are different generating a variance value (V_D). This is then compared with the expected variance value (V_e) when loci conform to linkage equilibrium (occurs when alleles assort independently) (Haubold & Hudson, 2000). The null hypothesis stating that $V_D = V_e$ is tested. To test for

significance, resampling of the alleles without replacement is performed by Monte Carlo simulation.

The microsatellites used in this study have been ascertained and validated for use in previous studies (Anderson *et al*., 1999; Anderson *et al*., 2000; Greenhouse, *et al*., 2006; Tsumori *et al*., 2011). The test of linkage disequilibrium was still performed in this study to determine whether the alleles are independently assorted or not. Only the genotypes obtained from the major clones were used. This is because including the minor clones would skew the data toward a high observed LD as a result of high genetic relatedness between the clonal genotypes. For this test, clonal genotypes obtained from Cameroon were not included as they comprised few samples.

5.3. Results

5.3.1. Identification of genotypes

A total of 13 SNPs were used and a mixed clone infection confirmed using the criteria described in **Chapter 4** and in **Section 5.2.1.1** to detect the clones represented at different proportions in a multiclonal infection.

All isolates were also analysed using microsatellites (Culleton R, Nagasaki University) which allowed the identification of additional mixed clone infections that were not otherwise detected by the SNPs. Of the 54 analysed isolates, 6 were identified as polyclonal by microsatellites only, 7 by Pyrosequencing^{TM} only, while 17 isolates were identified as polyclonal by both, the remaining 24 isolates were identified by both methods as single clones (**see Appendix F**). Altogether, following

identification of major and minor clones, 84 clonal genotypes were recovered and subsequently analysed using the methods already described in **Chapter 4** (**Table 5.1**).

Table 5.1 Summary of the clonal genotypes that could be identified from the 54 isolates by Pyrosequencing™ and microsatellites shown for either single clones or mixed clones

	Single Clone infections		Double Clone infections		
Country	**Isolates**	**No. of Clone genotypes**	**Isolates**	**No. of Clone genotypes**	**Total no. of clone genotypes**
Cameroon	2	2	5	10	**12**
Mali	19	19	15	30	**49**
Kenya	3	3	10	20	**23**
Total	**24**	**24**	**30**	**60**	**84**

SNPs and Microsatellites loci having missing data were removed as FSTAT would not compute the required statistics. Therefore of the complete set of 8 microsatellite loci, the results discussed in this chapter comprise data obtained from 5 microsatellite loci. In Cameroon, the loci TA40 and TA43 did not amplify by PCR and had missing data. In Kenya, the loci TA40 and TA60 did not also amplify by PCR and had missing data. Since these loci amplified successfully in Mali, this could have been due to these assays being sensitive to low DNA quality or yield obtained from the Cameroon and Kenya isolates.

12 SNPs were used in the analysis as the assay at the Pf-01-393153 locus failed to amplify the Kenyan isolates. This is perhaps due to the degradation of primers. Re-amplification of this locus was however not done due to time constraints.

5.3.2. Genetic diversity estimated by allelic diversity

The allele frequencies were calculated per locus, per allele and per population (in this case, Cameroon, Mali and Kenya) by estimating the number of individuals analysed and determining the overall allele frequencies. Both the weighted by sample size and unweighted allele frequency values were computed. The weighting by sample sizes is performed to eliminate bias introduced due to different sample sizes of the populations being studied. This is because allele frequency differences can be as a result of the sampled population not being an adequate representation of the overall population. The results indicated that there was considerable observed polymorphism per locus analysed (see **Appendix H**).

Table 5.2 (a) Summary of total number of alleles sampled and (b) the estimated allelic richness controlling for variation in sample size per locus across country

(a)

SNPS					MICROSATELLITES				
Locus	**Cameroon**	**Mali**	**Kenya**	**Total**	**Locus**	**Cameroon**	**Mali**	**Kenya**	**Total**
Pf-01-101502	2	2	1	**3**	Poly-α	6	12	9	**15**
Pf-01-205304	3	4	3	**4**	TA81	6	7	6	**8**
Pf-01-411796	1	2	1	**2**	TA17	8	10	6	**16**
Pf-03-138-7	2	2	3	**3**	TA53	3	5	3	**6**
Pf-03-350224	3	4	3	**4**	PFPK2	6	9	5	**10**
Pf-03-613146	1	1	3	**3**					
Pf-03-842383	1	1	1	**1**					
Pf-03-967838	1	1	1	**1**					
Pf-10-658467	1	2	3	**3**					
Pf-10-992433	2	3	2	**3**					
Pf-14-279668	2	3	3	**3**					
Pf-14-1710196	3	2	2	**3**					

SNPS					MICROSATELLITES				
Locus	**Cameroon**	**Mali**	**Kenya**	**Total**	**Locus**	**Cameroon**	**Mali**	**Kenya**	**Total**
Pf-01-101502	2.000	1.915	1.000	1.998	Poly-α	6.000	10.118	8.355	10.651
Pf-01-205304	3.000	3.534	3.000	3.439	TA81	6.000	6.188	5.856	6.897
Pf-01-411796	1.000	1.432	1.000	1.266	TA17	8.000	7.53	5.687	9.049
Pf-03-138-7	2.000	2.000	2.999	2.717	TA53	3.000	3.983	2.964	4.481
Pf-03-350224	3.000	3.418	2.992	3.24	PFPK2	6.000	7.619	4.769	7.775
Pf-03-613146	1.000	1.000	2.992	2.405					
Pf-03-842383	1.000	1.000	1.000	1.000					
Pf-03-967838	1.000	1.000	1.000	1.000					
Pf-10-658467	1.000	1.904	2.91	2.317					
Pf-10-992433	2.000	2.447	2.000	2.272					
Pf-14-279668	2.000	2.797	2.999	2.937					
Pf-14-1710196	3.000	2.000	2.000	2.609					

The number of alleles per country and per locus was counted as shown in **Table 5.2a**. The SNPs data revealed a balanced representation of alleles across the 3 countries. However, the microsatellites data showed that Mali had the highest number of alleles compared to Kenya and Cameroon. Further, the mean number of alleles per locus or the allelic diversity also known as allelic richness was also calculated (See **Table 5.2b**).

Information from the SNPs in **Table 5.2b** reveals that loci Pf-01-411796, Pf-03-842383 & Pf-03-967838 were the least informative as they had an average allelic richness of ~1. These SNPs are therefore not polymorphic rendering them inappropriate in the delineation of *P. falciparum* clones occurring in mixed infections within the studied geographic regions.

The microsatellites in **Tables 5.2a & b** indicate that Mali had the highest level of genetic diversity as explained by its number of alleles and estimated levels of allelic richness in comparison to both Cameroon (fixed sample size) and Kenya. This is demonstrated in the loci Poly-α, Ta53 and PFPK2 whereas in Kenya only the Poly-α locus had a higher allelic diversity estimate compared to the other geographic regions.

5.3.3. Unbiased estimates of overall genetic diversity

Having shown the degree of informativeness for both SNPs and microsatellites, the level of genetic diversity per locus and per geographic location was estimated. Gene diversity, also known as heterozygosity (H), is the probability that two alleles sampled from a population (to make a diploid genotype) are not identical. Thus gene diversity values of 0 mean that alleles are identical (no polymorphism) whereas a gene diversity value of 1 means that all alleles within a sampled population are not identical (highly polymorphic).

Genetic diversity values obtained from SNPs (**Table 5.3**) are low overall with Kenya having the highest estimate. Loci Pf-01-411796, Pf-03-842383 & Pf-03-967838 had insignificant levels of genetic diversity observed in them concordant with their low levels of allelic richness as in **Table 5.2b**.

Conversely, the results in **Table 5.3** indicate extensive genetic heterogeneity across the microsatellites loci analysed among the countries and overall. Only the locus TA53 had lower levels of gene diversity overall, as is clearly distinct in **Table 5.3** which is also explained by its low allelic richness values (**Table 5.2b**) rendering it

less informative compared to the other loci. Cameroon had the highest overall expected heterozygosity values indicative of high levels of polymorphism followed by Mali and then Kenya (**Table 5.3**).

The genetic differentiation estimate obtained by averaging multiple loci, G_{ST}' was estimated and is also given in **Table 5.3**. The G_{ST}' estimates obtained per locus in **Table 5.3** demonstrate the need to carry out a multilocus average estimation of genetic differentiation as different loci exhibit different values. The SNPs locus Pf-03-613146 had the highest G_{ST}' (0.0228) value relative to the other loci. Pf-14-279668 and Pf-14-1710196 also had relatively high G_{ST}' values i.e. 0.117 & 0.1 respectively. In this analysis, the microsatellites locus TA81 had the highest G_{ST}' value whereas Poly-α had the lowest.

Wright, 1978 (cited in http://www.library.auckland.ac.nz/subject-guides/bio/pdfs/733Pop-g-stats2.pdf) suggests the following qualitative guidelines for the interpretation of overall F_{ST} or G_{ST}' in this case (estimated from allozyme data)

- 0.0 to 0.05 may be considered as indicating little genetic differentiation
- 0.05 to 0.15 indicates moderate genetic differentiation
- 0.15 to 0.25 indicates great genetic differentiation
-above 0.25 indicate very great genetic differentiation

Given the guideline above, very little genetic differentiation was observed between the countries as revealed in **Table 5.3**.

Table 5.3 Gene diversity (unbiased estimates) or expected heterozygosity (Hexp) per locus, per subpopulation (in this case the countries: Cameroon, Mali and Kenya) and overall for both the SNPs and microsatellites data. The summary of Nei's estimation of heterozygosity as expected heterozygosity per subpopulation (Hexp (s)), total expected heterozygosity (Hexp (t´)) and the F_{ST} estimator $G_{ST'}$ independent of sample size is also given.

SNPS							**MICROSATELLITES**						
	Genetic Diversity							**Genetic Diversity**					
Locus	**Cameroon**	**Mali**	**Kenya**	**Hexp (s)**	**Hexp (t´)**	**GST´**	**Locus**	**Cameroon**	**Mali**	**Kenya**	**Hexp (s)**	**Hexp (t´)**	**GST´**
Pf-01-101502	0.167	0.159	0	0.108	0.11	0.016	Poly-α	0.848	0.921	0.887	0.886	0.908	0.024
Pf-01-205304	0.667	0.388	0.664	0.57	0.592	0.037	TA81	0.879	0.791	0.775	0.814	0.871	0.066
Pf-01-411796	0	0.041	0	0.014	0.014	0	TA17	0.939	0.833	0.802	0.857	0.903	0.051
Pf-03-138-7	0.303	0.463	0.648	0.474	0.503	0.058	TA53	0.545	0.481	0.567	0.53	0.554	0.043
Pf-03-350224	0.53	0.64	0.632	0.602	0.608	0.01	PFPK2	0.894	0.856	0.779	0.842	0.87	0.032
Pf-03-613146	0	0	0.534	0.179	0.232	0.228							
Pf-03-842383	0	0	0	0	0	0							
Pf-03-967838	0	0	0	0	0	0							
Pf-10-658467	0	0.153	0.316	0.159	0.163	0.029							
Pf-10-992433	0.409	0.509	0.522	0.481	0.484	0.006							
Pf-14-279668	0.167	0.321	0.632	0.376	0.426	0.117							
Pf-14-1710196	0.621	0.493	0.514	0.541	0.601	0.1							
Overall total	**0.239**	**0.264**	**0.371**	**0.292**	**0.311**	**0.061**	**Overall total**	**0.821**	**0.7764**	**0.762**	**0.786**	**0.821**	**0.043**

Following calculations by FSTAT using microsatellites data only, the R_{ST} value obtained was 0.0318. This R_{ST} value consistent with the step-wise mutation model (SMM) is less than, but not significantly different from the estimated F_{ST} value (G_{ST}´or θ value) obtained using the Infinite alleles model (IAM) in **Table 5.3**. (Conway *et al*., 2001) also observed non-significant differences between F_{ST} and R_{ST} values in a study conducted in isolates obtained within Africa although significant differences were observed between intercontinental estimates. In addition, F_{ST} estimates obtained from SNPs only (using IAM) had a value of 0.061 (**Table 5.3**) which is also not different from that observed with the microsatellites data. This indicates that the SMM model can be ignored especially when studying short-term parasite generations where mutations have little influence on their genetic diversity.

5.3.4. Population structure

To test for population structure between the countries, the clonal genotypes were uploaded to the software FSTAT in separate files i.e. either for SNPs or microsatellites. The results of the statistics are summarised in **Table 5.4**. Microsatellites analyses had statistically significant levels of genetic differentiation between Cameroon & Mali, Mali & Kenya and Cameroon & Kenya. However, only the Mali & Kenya country comparisons were significant with the SNPs analysis (**Table 5.4**). Based on the microsatellites data, the null hypothesis stating that the populations are not genetically differentiated was therefore rejected.

Table 5.4 Summary statistics from the SNPs and microsatellites analysis for the computed genetic differentiation obtained by pairwise comparisons of sequences between: Countries (Cameroon, Mali & Kenya), Mali clusters (1, 2 & 3) and Mali aureoles (1, 2 & 3) comparisons. The clusters from the Mali study site are obtained from the partitioning described in **Figure 5.1**. The aureoles comprise the boundaries defined in the map provided also in **Figure 5.1**. Statistically significant values at P <0.05 are given in red font colour.

SNPS			**Microsatellites**		
Country comparison					
	Mali	Kenya		Mali	Kenya
Cameroon	0.0563	0.0734	Cameroon	0.0355	0.058
Mali		0.0485	Mali		0.037
Clusters comparison					
	Cluster 2	Cluster 3		Cluster 2	Cluster 3
Cluster 1	0.0695	-0.0422	Cluster 1	0.0641	0.0581
Cluster 2		-0.0065	Cluster 2		0.0055
Aureoles comparison					
	Aureole 2	Aureole 3		Aureole 2	Aureole 3
Aureole 1	-0.0363	0.0139	Aureole 1	0.0311	0.0876
Aureole 2		0.0083	Aureole 2		0.0527

The clonal genotypes obtained from Mali were subjected to further geographic partitioning and the genetic differentiation estimates shown in the clusters comparisons section of **Table 5.4** computed. This was done to investigate whether genetically related malaria parasites occur within small spatial scales and hence test the hypothesis that parasites occurring within small spatial settings are highly structured and comprise closely mating parasite subpopulations.

Mali was chosen as it had the highest sample representation and also because the exact geographic locations of its samples are known. In this study site the clonal genotypes were grouped according to their point of collection (**see Figure 5.1**).

The identified clusters 1, 2 & 3 in **Figure 5.1** were defined in an attempt to encompass as many homesteads as possible for adequate sample representation in the intended analysis. The distances obtained between the houses and described in **Table 5.5** were determined manually using the scale provided aided by the use of a 30cm ruler.

Table 5.5 Information on the partitioning performed on the Kolle region, Mali. The geographic distance shown is between the two farthest sampled households per cluster. The number of clonal genotypes analysed per cluster is also shown.

Partitioning (Cluster)	**Maximum geographic distance (based on largest separation distance of households)**	**No. of analysed clonal genotypes**
1	300m	10
2	250m	16
3	300m	15

The results in **Table 5.4** indicate that no statistically significant level of genetic differentiation was detected with these partitions ($P < 0.05$).

Further investigation of population structure was performed using the already defined boundaries in the form of aureoles within the Mali study site. This demarcation was established by the principal investigators in charge of the project at this study site for easy sampling of individuals in the households (see **Figure 5.1**). Only aureoles 1, 2 & 3 were included in this analysis as they had adequate sample representation. Thus, 10 clonal genotypes from aureole 1, 28 from aureole 2 and, 10 from aureole 3 giving a total of 48 clonal genotypes were included in this analysis. With this analysis, there was genetic differentiation between the clonal genotypes obtained from Aureoles 1 & 3 (**Table 5.4**).

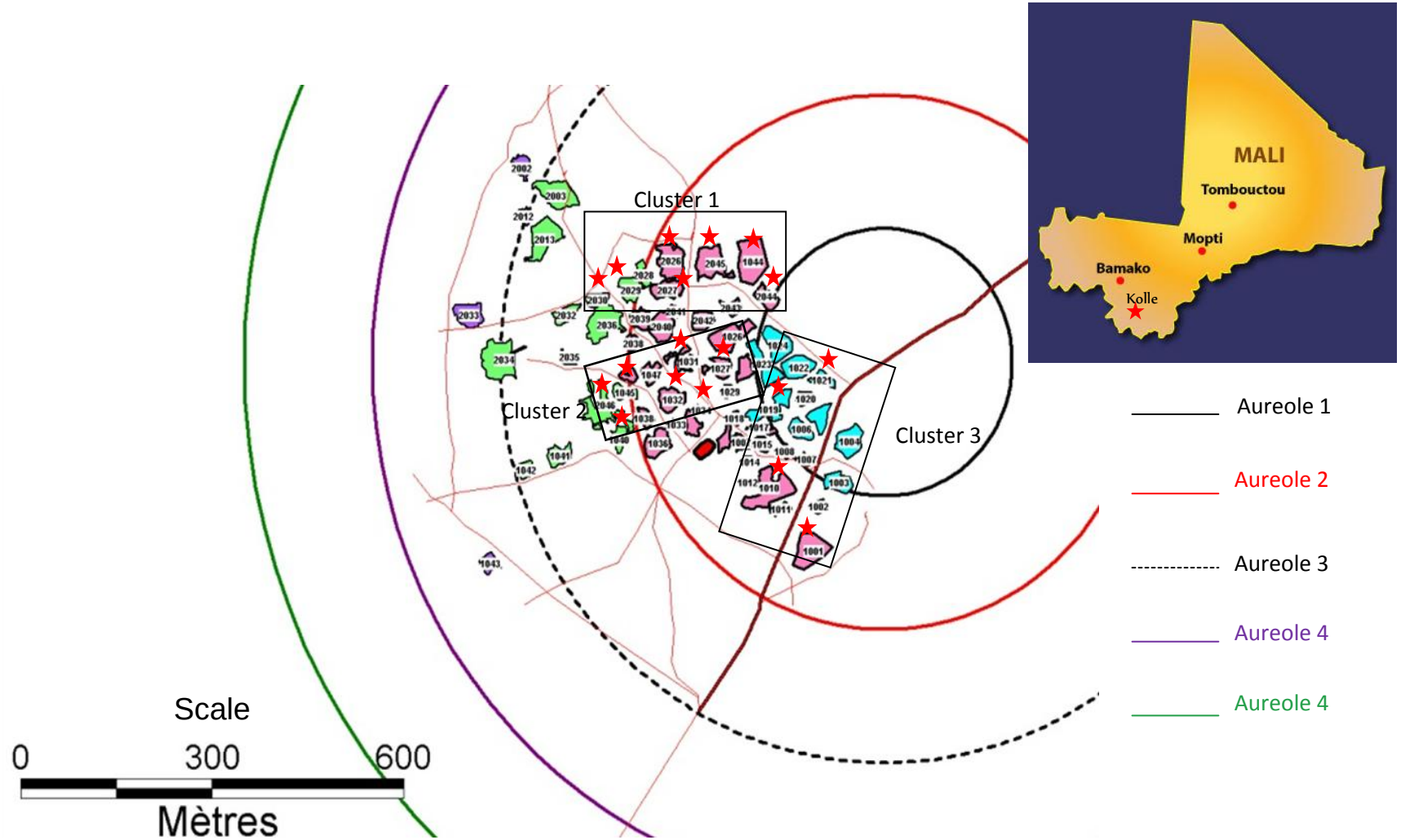


Figure 5.1 Map of Kolle, Mali showing the households partitioned into aureoles (each 200m) and also the further partitioning performed in an attempt to determine genetic differentiation within small geographic settings. River Kolle, the breeding site, is the brown line cutting across the aureoles. The coloured sections i.e. blue, pink, green or purple comprise the households or families. The partitions (clusters 1-3) shown by the rectangles are as described in **Table 5.5**. The red stars indicate the sampled households from which the 41 clonal genotypes were obtained. Map of Mali showing the location of Kolle region is also given.

Additional analysis of clonal genotypes obtained from Rusinga Island, Kenya (see map in **Chapter 2**) was carried out to determine levels of genetic differentiation. The locations included in the analysis are Kaswanga and Kamasengre. This is because they had adequate sample representation. However, the actual households or physical locations from which the samples were obtained are unknown as collections were made in the local health centres. These two regions are separated by approximately 5 km. 11 clonal genotypes from Kaswanga and 10 clonal genotypes were included in the analysis. Interestingly, this analysis detected statistically significant genetic differentiation with the SNPs (F_{ST} = 0.0472) and microsatellites (F_{ST} = 0.15).

5.3.5. Linkage disequlibrium (LD)

The parameter I_A^S is used to determine the magnitude of the observed LD. The results (**Table 5.6**) obtained indicate that no significant levels of LD were observed. The low values of I_A^S are an obvious indication of this and the insignificant P values also corroborate the results.

Table 5.6 Analysis of LD at a country level given as the overall value across the loci analysed.

Geographic Region	**No. of identified genotypes**	**Index of association (I_A^S)**	**Statistical significance (P value)**
Mali	14	-0.015	0.7
Kenya	11	0.019	0.356

5.4. Discussion

It is important to note that the purpose of this study is not to review the performance of genetic markers in population genetic studies but to establish appropriate markers for the study of the interaction of malaria parasites within small geographic settings. However since both SNPs and microsatellites were used in this study, it is pertinent to review the findings obtained in this study and their implications for future planned studies of this nature.

5.4.1. Genetic diversity

Measurements of genetic diversity involve the estimation of divergence in DNA sequences. Malaria parasite lineages over time are under various forces at a genome level such as genetic recombination, mutations or natural selection which are responsible for the generation of nucleotide divergence. In addition, their interaction at a population level (their gene flow patterns) predisposes them to genetic drift or inbreeding due to mating constraints imposed by the natural phenomenon of isolation by distance leading to population structuring. Since it is costly to analyse whole genomes for extrapolation of genetic diversity, markers obtained from multiple loci are used. It is pertinent that sufficient genetic markers are utilized to ensure they adequately represent the existing genome wide information.

SNPs are considered less informative in the estimation of genetic diversity as they are biallelic and have low mutation rate compared to the highly variable microsatellites (Liu *et al*., 2005). This is evidenced by the differences in number of alleles and allelic richness observed with either the SNPs or microsatellites (**Table**

5.2 a&b). Even though only a few microsatellites were used in this study, they were indeed more informative compared to the SNPs as they displayed greater allelic diversity and richness (**Table 5.2 a&b**). This means that for accurate estimation of nucleotide sequence diversity, numerous SNPs would be required. This setback is however mitigated by the relative abundance of SNPs in the genome easing their identification and use in genetic studies (Morin *et al*., 2004). On the other hand, although abundant too, the development of microsatellites' assays is more time consuming and laborious (Haasl & Payseur, 2011).

The SNPs analysis revealed that Kenya had the highest levels of genetic diversity followed by Mali and Cameroon had the lowest (**Table 5.3**). In contrast, the microsatellites estimate of heterozygosity (**Table 5.3**) revealed that Cameroon was the most genetically diverse, then Mali and finally Kenya. The disparity in estimates of genetic diversity observed in Cameroon, Kenya and Mali using either microsatellites or SNPs is not a new phenomenon and has been observed and investigated previously by the comparison of AFLPs and microsatellites (Mariette *et al*., 2001; Mariette *et al*., 2002) or microsatellites and SNPs (Väli *et al*., 2008). The explanation for this is either the use of few genetic markers that fail to capture the heterogeneity of genomes or the analysis of a recently created population (Mariette *et al*., 2002). In this study, it is possible that the few genetic markers used failed to clearly capture the overall genetic diversity present in malaria parasites for each of the study sites analysed. Additionally, in a recently created population, such disparities are expected. This is because SNPs take longer to mutate and capture the existing genetic divergence within populations. Thus, the low within genomic heterogeneity fails to capture the within population genetic diversity. On the other

hand the high within genomic diversity demonstrated by microsatellites overestimates the within population heterogeneity.

In the estimation of allele frequencies, studies recommend the use of the predominant clone only to minimise ascertainment bias (see Anderson *et al*., 2000 using microsatellites; Anderson *et al*., 2005 using SNPs). Ascertainment bias in the estimation of genetic diversity arises from the overestimation of rare allele frequencies or underestimation of common allele frequencies (Ellegren, 2004). However, in this study, all the identified clonal genotypes were used in the estimation of allele frequencies as use of the major clone only would lead to data losses. Since the software FSTAT uses allele frequencies to estimate F_{ST} values, the existence of bias in the obtained results cannot be negated.

The high levels of gene diversity or expected heterozygosity [Hexp (s)] observed per locus and per geographic location (**Table 5.3**) using microsatellites represent a typical scenario in areas of high malaria transmission such as in Africa. Similar findings were observed using microsatellites by Anderson *et al*., (2000) in Uganda, Congo and Zimbabwe; by Tsumori *et al*., (2011) in Congo and Mobegi *et al*., (2012) across West Africa. On the other hand, the low values of genetic diversity observed with the SNPs data (**Table 5.3**) are because a few of them were used in this study rendering it difficult for the SNPs to effectively estimate the genetic divergence between the countries. In their review, Haysl & Payseur (2011) suggest that increasing the number of SNPs in the detection of population divergence may give estimates comparable to microsatellites.

The results obtained in this study indicate that the choice of genetic markers is very important in the interpretation of genetic diversity estimates. Microsatellites are recommended for the estimation of genetic diversity in recent populations (Avise, 2004). In addition, a few of them are able to detect levels of population divergence as observed in this study. Since this study is concerned with recent genetic events, microsatellites should suffice. However, the enormous prospects of using the highly abundant SNPs in the *P. falciparum* genome coupled with the increased progress in sequencing technology (Mardis, 2008) for the detection of within genomic or within population diversity is not to be downplayed. For example, Manske *et al*., 2012 has used the newly developed Illumina genetic analyser to identify about 86,000 SNPs appropriate for population diversity measures. Nonetheless, it is pertinent that sufficient genetic markers are used to provide accurate measures of existing population divergence.

5.4.2. Linkage disequilibrium

It is noteworthy that LD can occur at the genomic level either due to natural selection or in regions of low recombination rate or if alleles are in close proximity in the genome such that they are linked. It can also occur at the population level as a result of inbreeding or genetic drift. In *P. falciparum*, significant levels of LD have been observed in regions of low malaria transmission such as Bolivia (Anderson *et al*., 2000), Malaysia (Anthony *et al*., 2005), Thailand (Pumpaibool *et al*., 2009) and Phillipines (Iwagami *et al*., 2009) but not in high transmission intensity regions such as is observed in Uganda and Congo (Anderson *et al*., 2000 and in West Africa

(Mobegi *et al*., 2012). Thus the lack of LD observed in this study (**Table 5.6**) signifies that indeed the microsatellites are neutral and represent independent loci within the *P. falciparum* genome. In addition, these observations corroborate with the existing evidence of no LD when dealing with isolates obtained from regions of high malaria transmission (Anderson *et al*., 2000; Mobegi *et al*., 2012).

5.4.3. Genetic differentiation

This study was devised to infer the gene flow patterns of malaria parasites within small geographic settings (<1km) and thereby determine the size (in terms of geographic distance) of a closely mating population having high levels of genetic relatedness. According to the isolation by distance theory, closely mating parasites are indeed expected to be genetically related. However in *P. falciparum*, within a focus of malaria transmission, the geographic size of a cluster of closely mating populations (constituting a deme) is unknown.

Closely mating malaria parasites are highly inbred as they are isolated geographically with limited migration between subpopulations. This leads to reductions in heterozygosity (genetic diversity) due to non-random mating. The parameter genetic differentiation (F_{ST}) is used to measure this. If the geographic locations of the parasite isolates are known, then F_{ST} values can be determined and the points at which no statistically significant F_{ST} values are detected indicate the limit of migration of the parasites and offer an estimate of how widespread a single subpopulation is.

5.4.3.1. Overall F_{ST} values

The overall G_{ST}' (F_{ST}) value of 0.061 using SNPs and 0.043 with microsatellites (**Table 5.3**) obtained by pairwise comparison of genotypes across the different countries in Africa is low and consistent with what was detected by Anderson *et al*., (2000) in Uganda and Congo and recently by Mobegi *et al*., (2012) within and among West African countries. These low values of F_{ST} are typically observed in African countries separated by geographic distances >1000km. Higher F_{ST} values (> 0.05) have been observed in regions separated by the same geographic distances in South America and Asia (Anderson *et al*., 2000), in the Brazillian Amazon region (Machado *et al*., 2004), and in Papua New Guinea (Schultz *et al*., 2010). This makes the detection of genetic structure in regions of smaller geographic distances within Africa difficult as very low genetic differentiation is in existence across regions separated by very large geographic distances. For example Tsumori *et al*., (2011) in regions separated by approximately 8km detected an F_{ST} value of 0.004 indicative of panmixia.

At a continental level, low gene flow is expected between malaria parasites hence the observed statistically significant levels of population structure in the between country comparisons. However, in Africa low values of genetic differentiation such as was observed here have been previously recorded (Anderson *et al*., 2000). These low values indicate that most of the genetic variation in Africa is distributed within subpopulations as opposed to among subpopulations. It is pertinent to note that in Africa, a spectrum of genetic differentiation values is observed due to the varying malaria transmission patterns across the continent. For example in a study conducted in low malaria transmission urban and rural sites of West Africa i.e. Djibouti,

Niamey, Senegal & Dakar, high F_{ST} values were reported (>0.1) not indicative of panmixia (Bogreau *et al*., 2006). In another study conducted across West African countries of varying malaria prevalence detected low levels of F_{ST} (~0.03) (Mobegi *et al*., 2012) concordant with what was observed by Anderson *et al*., 2000 in Africa. These studies show that although Africa is considered to be a high malaria transmission zone, determining the population genetic profiles within sites of interest is important.

5.4.3.2. Between country comparisons

The results indicated low but significant levels of structuring with the data obtained from microsatellites (**Table 5.4**). Such low levels of genetic differentiation were also observed between Mali and Burkina Faso (both West African countries) (Campino *et al*., 2011). The SNPs data indicted that only between country comparison of Mali & Kenya yielded statistically significant F_{ST} values (**Table 5.4**). Comparable levels of genetic differentiation across Africa were also observed with SNP analysis of the highly polymorphic *Plasmodium falciparum* sarco/endoplasmic reticulum calcium-ATPase (*pf*SERCA) gene, implicated but not yet proven to play a role in the development of parasite drug resistance to the antimalarial, artemisinin (Jambou *et al*., 2010).

In Mali, partitioning by aureoles revealed that individuals residing in tightly clutered households (**see Figure 5.1**) harboured parasites that were closely related genetically. This was demonstrated by the statistically significant levels of genetic differentiation obtained from the comparisons between aureoles 1 & 3 and not between aureoles 1

& 2 (**Table 5.4**). A close inspection of the study site in **Figure 5.1** reveals that the breeding site cuts across the households in these aureoles such that most individuals reside in close proximity to the river Kolle.

These results are not surprising as the existing literature on the transmission patterns of *P. falciparum* indicate that regions closest to mosquito breeding grounds experience the highest levels of malaria transmission (Ghebreyesus *et al*., 1999; (Thomas & Lindsay, 2000; Staedke *et al*., 2003; van der Hoek *et al*., 2003; Midega *et al*., 2012). This is because of high mosquito abundance around breeding sites such that individuals residing in the vicinity of these areas experience the highest mosquito bites accompanied by parasite inoculations. Since the mosquitoes or the people (harbouring the parasites) exhibit limited migration, the parasite population around these sites is expected to be highly inbred with low genetic variation. However, it is noteworthy that intepration of these results should be taken with caution as only a small data set was used in this analysis. Also, few genetic markers were used in this study and it is possible that they lacked the power to accurately detect the levels of genetic variation within the Mali study site.

It is noteworthy that the demarcation by aureoles used in the analysis of the Mali study site is not ideal for this kind of study. This is because households at the border of the aureoles are within close geographic distance obscuring patterns of genetic diversity in relation to physical distance. The Study site in Mali offers good prospects in the study of malaria parasites' transmission patterns within short geographic distances. However, adequate sampling is required for accurate estimations to be obtained.

Even though the statistically significant F_{ST} value obtained by the comparison of isolates obtained from Kaswanga and Kamasengre in Rusinga Island, Kenya from the SNPs data (0.0472) seems reasonable and accurate compared to the one obtained using microsatellites (0.15), since only a few isolates were analysed. However, there is also a high probability that these values are artifactual as only a small sample set was included in this analysis.

In conclusion, the hypothesis that parasites occurring within small spatial settings are highly structured and comprise closely mating parasite subpopulations cannot be fully validated. In the next chapter, the complete data set obtained from the SNPs and microsatellites will be combined for genetic distance based analyses to elucidate patterns of malaria parasites distribution in space as explained by their genotypes.

5.4.4. Conclusion

It is noteworthy that the SNP and microsatellites data analysis in this chapter reveals that within a distance of approximately 400m from the breeding site, there is significant genetic structuring of the malaria parasites. This means that in this study site, based on the results obtained, successful malaria transmission is maintained within a radius of approximately 400m from the breeding site. This however, does not offer conclusive evidence on population structuring within the Mali study site. The low number of SNPs used in this study did not have the detection power to reveal further genetic differentiation if it exists in the Mali region. Additional genetic markers particularly in the Mali study site are therefore required to provide a clearer picture.

The information obtained from Kenya was also inadequate as the actual geographic location of the samples is unknown. A proper study design to investigate the transmission patterns in Rusinga Island is required to trace the genetic relatedness of malaria parasites in this study site.

6. GENETIC DISTANCE BASED ANALYSIS OF THE *P. FALCIPARUM* ISOLATES

6.1. Introduction

An important question population geneticists ask addresses how individuals interact in space and how this affects their degree of genetic relatedness or diversity. Genetic diversity is relevant for maintenance of fitness in natural populations. For example in captive breeding programs for either production of salmon (reviewed by Naylor *et al*., 2005; Hindar *et al*., 2006; Fraser, 2008) or protection of endangered species (reviewed by (Matocq & Villablanca, 2001; Williams & Hoffman, 2009) there is danger of inbreeding and loss of genetic variability. In the wake of advanced Agricultural technology, there is concern over the consequences imparted by genetic introgression i.e. the transfer of genes from a genetically modified crop to their free living wild counterparts (reviewed by Stewart *et al*., 2003). It has been implied that this may lead to fitness loss or extinction of wild plants brought about by competitive advantage of novel genes that may homogenize the existing genotypes (Mooney & Cleland, 2001).

In health research, understanding the interactions of disease causing pathogens is essential to monitor their transmission patterns including the emergence and spread of genes responsible for increasing their virulence or drug resistance. For example a recent review on opportunistic bacterial pathogens i.e. bacteria that are otherwise harmless but cause disease in immune-compromised humans, appreciates that these bacteria develop virulence or antibiotic drug resistance in their primary hosts such as

plants. Their development of adaptive mechanisms in these environments renders them capable of evading the human hosts' immune response and perhaps introducing novel virulence factors to the pre-existing mutualistic or pathogenic bacteria (Aujoulat *et al*., 2012).

In the vector-borne disease, *P. falciparum* malaria, transmission of these parasites from person to person by the female *Anopheles* mosquito is vital to understanding their interaction in space. It is in the mosquito that genetically distinct malaria parasites undergo genetic recombination giving rise to novel genotypes. However, the probability for the mosquito to pick up two genetically distinct parasite clones depends on parasite genetic diversity within infected human hosts. Thus in localities, the cycle of repeated parasite inoculations by the mosquito from one individual to another is constantly replayed. The assumption would be that malaria parasites obtained from people located in close neighbourhoods bear parasites that are closely related genetically in comparison to individuals that are far apart.

To understand the concept of gene flow patterns in malaria parasite populations, the impending question is: are the parasites within a population randomly mating or do they constitute discrete subpopulations with limited gene flow between them? To address this, Sewall Wright in 1943 developed the isolation by distance model which states that the genetic relatedness of individuals reduces with increase in geographic distance between them. It is on this premise that genetic distance based analyses are founded.

In this final chapter, the complete data set combining the SNP and microsatellite genetic loci will be analysed. The gene flow patterns of these parasites will then be

elucidated to identify if they comprise discrete clusters of closely mating parasites and at what geographic distances, or if they comprise randomly mating parasites that are not genetically structured. Using genetic distance based methods the isolation by distance model will be tested on *P. falciparum* isolates collected from the different geographic regions.

6.1.1. Synopsis of chapter

This chapter highlights the main theme of this thesis – the genetic relatedness of malaria parasites in space. The data obtained from both SNPs and microsatellites were combined to perform genetic distance based analyses in contrast to the frequency based analyses. The difference between the two is that the latter comprises extrapolation of genetic variation from differences in allele frequencies whereas the former utilizes genetic distance matrices obtained from pairwise comparison of DNA sequences to infer genetic variation or similarity in space (Peakall and Smouse, 2006; 2012). In addition, these tests do not conform to mutation based models allowing data obtained from both SNPs and microsatellites to be combined for analysis.

In this chapter, pairwise comparisons of the obtained genotypes were obtained and genetic distances (GD) computed. A GD matrix was obtained and used for subsequent analysis to determine the relationship between GD and geographic distances of the *P. falciparum* populations in question.

The overall genetic variation among and between the parasite populations obtained from different geographic settings were determined. Further to this, the correlation between GD and geographic distance over distances >1km and beyond was investigated. Finally, a spatial autocorrelation analysis was performed to investigate the existence of discrete clusters of malaria parasites within small spatial scales (<1km).

Since the geographic coordinates of the isolates obtained from Mali are known, a separate analysis was also carried out with these isolates obtained from shorter geographic distances (<1km).

6.1.2. Objective of chapter

As already stated in **Chapter 1**, this thesis constitutes a feasibility study to investigate whether the genetic relatedness of malaria parasites can be used to trace their transmission in space. In **Chapter 5**, the identified parasite genotypes obtained using both SNPs and microsatellites were analysed for evidence of population structure within predefined geographic boundaries. The results seemed to indicate that there are clusters of malaria parasites occuring within a distance of ~400m. In this chapter, further exploration of these study sites to reveal distinct clusters of closely mating parasites and perhaps reveal how malaria parasites are transmitted from person to person will be carried out.

6.1.3. Hypotheses of study

The null hypothesis (H_0) is that malaria parasites are randomly mating in space and are not genetically structured. The alternative hypothesis (H_1) is the isolation by distance hypothesis which states that closely related parasites occur within short separation distances of each other in space and that their level of genetic relatedness dissipates with increasing geographic distance.

6.2. Materials and methods

All the clonal genotypes obtained using SNPs and microsatellites were analysed in this chapter (**Appendix F**). The genetic based analyses were perfomed using the Microsoft Excel, add-in software GenAlEx 6.5 (Peakall and Smouse 2006; 2012). The following tests were performed: Analysis of molecular variance (AMOVA) to determine the genetic variation within and among the parasite populations. Mantel test, a statistical tool used to define population structure of the parasites over large geographic distances (>1km). It provides measurements of statistical significance to the correlation of matrices and in this study it will be used on genetic distance (as a measure of genetic variation) and geographic distance matrices. Finally, the spatial autocorrelation test in the identification of closely mating malaria parasite clusters within short geographic distances (<1km).

6.2.1. Genotyping isolates

All the *P. falciparum* clonal genotypes obtained per isolate were included in the analysis. Thus, the numerically coded data comprised a total of 84 clonal genotypes (12 from Cameroon, 49 from Mali and 23 from Kenya). In addition, 17 loci (5 microsatellites and 12 SNPs) were uploaded to the software. The alleles per locus for the SNPs data were assigned the numerals 1 =A, 2 = C, 3 = G and 4 = T. For the microsatellites, the alleles per locus were assigned the numbers between 1 for the lowest size allele observed across all loci and 15 for the highest observed allele size. Loci containing missing data were excluded. The latitude and longitude co-ordinates for each of the clonal genotypes were uploaded in separate columns labeled 'x' and 'y' on the same sheet containing the data set. Since the actual geographic coordinates of the parasite isolates obtained from Cameroon and Kenya were unknown, the coordinates given comprised rough estimates of the villages from where the isolates were collected.

49 *P. falciparum* clonal genotypes obtained from Mali only were also uploaded in a separate GenAlEx worksheet as the geographic coordinates were known. The data were partitioned into parasite populations from each of the aureoles (1, 2 & 3) described in the map of Mali (**Figure 5.1**). Thus 12 were from Aureole 1, 26 from Aureole 2 and 11 from Aureole 3.

6.2.2. Calculating distance matrices between individuals

A haploid genetic distance matrix was first generated in GenAlEx (both for the whole data set and the data set obtained from Mali) by performing an individual by individual pairwise comparison. Genotypes that are identical generate a GD of 0 whereas completely different genotypes generate a GD of 1. The sum of GDs across all loci was computed so that for a given pair of genotypes, the GD was equal to the tally of differences between the two. In addition, two separate geographic distance matrices were generated using the actual geographic distances and the log transformed geographic distances (log 1+x).

6.2.3. AMOVA analysis

The hierarchical partitioning of genetic variation among and within populations is typically investigated in population genetic studies. The parameters used include F_{ST} (Wright, 1931; 1951 cited in Hamilton, 2009) and its analogues: G_{ST} (Nei, 1987), θ (Weir & Cockerham, 1984), R_{ST} (Slatkin, 1995), Φ_{ST} (Excoffier *et al*., 1992). The problem with these measures is their dependency on within population genetic variation which confounds observed levels of genetic differentiation such that high levels of this variation lead to low F_{ST} values (Hendrick, 2005). It is on this premise that modifications of the Φ_{ST} parameter (Excoffier *et al*., 1992) have been developed to allow the analysis of genetic markers having different mutation rates such as SNPs and microsatellites. The AMOVA framework (Meirmans, 2006; Peakall and Smouse,

2006; 2012) allows the direct measure of population differentiation on any form of genetic data and precludes the need for determining allele frequencies.

Using the generated GD matrix, an AMOVA was performed which tests the degree of genetic variation within and among subpopulations (genetic differentiation) as given by the parameter Φ_{PT} (Excoffier, 1992) which is analogous to F_{ST} (Wright, 1931; 1951 cited in Hamilton, 2009). The formulae used in this analysis are summarised in **Appendix G**.

The parameter Φ_{PT} falls within the range of 0-1 and its tendency to 1 is indicative of increasing genetic differentiation. This parameter gives an estimate of the proportion of variance present within populations relative to the total variance.

The AMOVA test is performed under the H_0 that there is no genetic differentiation among subpopulations ($\Phi_{PT} = 0$) such that the data set comprises randomly mating subpopulations with little genetic differences between them unless those generated by chance due to sampling effects. Statistical testing was performed using 1000 permutations where the data was reshuffled or randomized multiple times to determine if the observed estimate is greater than that expected by chance with a P value of 0.05.

6.2.4. Mantel test

In natural populations, individuals that are separated by short geographic distances have a higher chance of mating than those separated by large distances. This is known as the isolation by distance phenomenon originally postulated by Wright,

(1943). The model operates under the assumption that there is a positive correlation between genetic distance and geographic distance.

The Mantel test (Mantel, 1967) enables the statistical comparison of the GD and geographic distance matrices. Thus for the two matrices (denoted by X & Y in **Appendix G**), a correlation coefficient value, R_{xy} is computed and its level of statistical significance defined using the H_0 that no relationship exists between the two matrices ($R_{xy} = 0$). R_{xy} spans within a range of +1 and -1. A null distribution is generated by holding the components of one matrix constant while randomizing the other one. The generated observed value is then compared with the permuted value and is considered significant if it is greater than is expected by chance to the permuted value. The Mantel test is used to examine patterns of population distribution over large geographic distances (>1km and beyond). Statistical significance was determined following 1000 randomisations at a P value of 0.05.

The geographic co-ordinates uploaded to GenAlEx were in the form of latitude and longitude in decimal degrees. The distance (D) was then computed in kilometres (km) (**Appendix G**).

6.2.5. Spatial autocorrelation analysis

The spatial autocorrelation analysis determines patterns of distribution in individual genotypes within short geographic distances (<1km). This test is ideal for detecting clusters of population structure within small study sites.

GenAlEx performs this test by a multivariate spatial autocorrelation approach (Smouse & Peakall (1999); Peakall *et al*., (2003); Double *et al*., (2005); Smouse *et al*., (2008) circumventing the classical allele-allele and locus-locus analyses which is subject to stochastic noise. Genetic and geographic distance matrices already generated for the Mantel test analysis will be used. The genetic distance matrix is converted to a covariance matrix which is compared to the geographic distance matrix. The correlation coefficient obtained (r) provides an estimate of the genetic relatedness of individuals falling within specified geographic distance classes with results output in the form of a correlogram. The value of r falls within a range of +1 and -1.

This test is performed with the null hypothesis (H_0) that genotypes are randomly distributed in space irrespective of their geographic separation (r = 0). The test for statistical significance is determined by random permutations and bootstrap estimates of r. The permutation test randomly shuffles individuals multiple times (performed 1000 times) among the geographic locations and generates an r permutation value (r_p) value under the H_0 (random distribution of individuals in space with no spatial structure). The r_p value is bound by upper and lower limits defined within a 95% confidence interval (two tailed t-test). If the calculated r value falls outside the bounds of this confidence limit, H_0 is rejected meaning there is spatial structuring.

Testing by bootstrapping allows confidence intervals to be estimated from pairwise comparisons within the predefined distance classes. The upper and lower limits of the observed r estimate are obtained by drawing with replacement at each distance class. 1000 bootstraps were performed and the bootstrap autocorrelation coefficient (r_{bs}) calculated for each distance class. Within a 95% confidence interval, if the

bootstrap confidence interval falls out of r = 0, the H_0 is rejected with significant spatial structure detected.

The distance classes were defined depending on the geographic distances of separation for each of the analysed data sets. Each distance class is represented by its own x matrix. If the samples represented in the pair wise comparison fall within the distance class, the elements have a value of 1 otherwise all other elements falling beyond this distance class are not represented and have a value of 0. The threshold of the distance classes is therefore set by adding one class above the maximum geographic distance of the data set.

6.3. Results

The analyses performed in this section aim to address the main question of this thesis – are malaria parasites occurring within short geographic distances (<1km) highly genetically related in contrast to those far apart (within distance of separation >1km)? To address this, the clonal genotypes obtained by SNP and microsatellites analysis were combined as the tests performed here are genetic/geographic distance based irrespective of the underlying evolutionary processes such as mutations occurring within the genetic markers of choice (Peakall & Smouse, 2012).

6.3.1. Analysis of molecular variance (AMOVA)

6.3.1.1. Overall AMOVA analysis

AMOVA was performed in the overall data set to determine the degree of genetic variation explained within or among sampled *P. falciparum* isolates. The combined *P.falciparum* genotypes were uploaded to GenAlEx and individual by individual pairwise comparisons performed. A genetic distance matrix was generated and used to compute the genetic variation observed within and among the populations as summarised in **Table 6.1** below.

Table 6.1 Summary of AMOVA analysis across the full data set

Source	Degrees of freedom	Sums of squares (SS)	Mean sums of squares (MS)	Estimated Variance	% variance
Among Populations	2	16.877	8.439	0.195	5%
Within Populations	81	308.765	3.812	3.812	95%
Total	83	325.643		4.007	100%

The results displayed in **Table 6.1** indicate that the majority of genetic variation in the overall *P. falciparum* isolates obtained from Cameroon, Kenya and Mali is explained within the populations (95% variance) as opposed to among the populations (5%).

In this analysis, the overall mean observed Φ_{PT} was 0.049 and it was statistically significant with a P (rand >= data) of 0.001 i.e. the P value obtained following randomisations. Further evidence to this, the frequency distribution of the observed

Φ_{PT} compared to the frequency of the random Φ_{PT} value obtained following 1000 random permutations was also higher and therefore statistically significant as shown in **Figure 6.1**. The H_0 stating that the data comprises randomly mating individuals that are not genetically differentiated was therefore rejected.

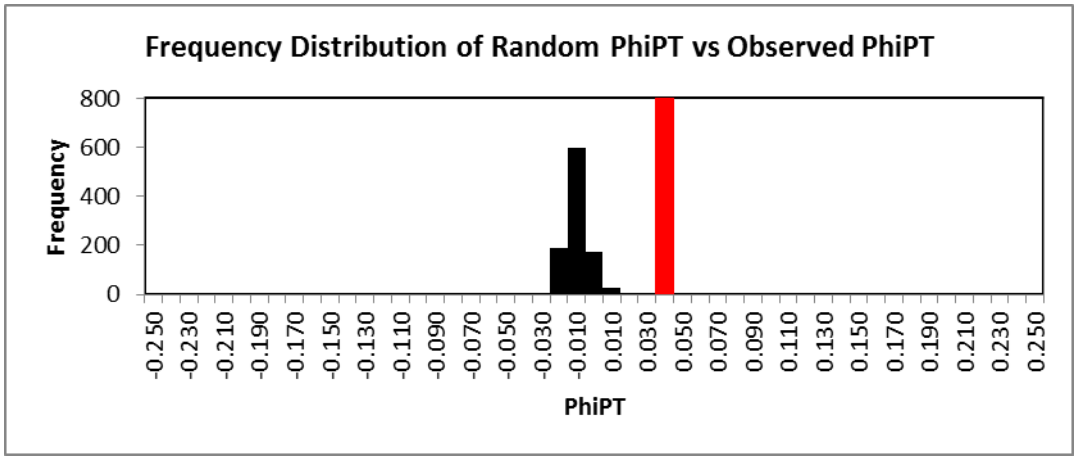


Figure 6.1 Frequency distribution of random Φ_{PT} (depicted in black) vs. observed (in red) following randomisation as a test for statistical significance in the overall analysis of clonal genotypes

Finally, country comparisons of the parasite populations were performed and the Φ_{PT} values obtained from the analysis are shown in **Table 6.2**. The results indicate that significant genetic differentiation was observed by comparisons between countries as shown by P (rand >= data).

Table 6.2 Country comparison of *P. falciparum* genotypes with the corresponding observed and permuted Φ_{PT} values shown.

Country comparison	**Distance (km)**	**Observed Φ_{PT}**	**P (rand >= data)**
Cameroon - Mali	2300	0.046	0.019
Cameroon - Kenya	2800	0.065	0.028
Mali - Kenya	5100	0.046	0.006

6.3.1.2. AMOVA test – Mali clonal genotypes

The AMOVA test was also performed in *P. falciparum* isolates obtained from Mali to evaluate the hierarchical partitioning of genetic variation among and between the aureoles. The results are shown in **Table 6.3**.

Table 6.3 Summary of the AMOVA analysis in Mali.

Source	Degrees of freedom (df)	Sums of squares (SS)	Mean sums of squares (MS)	Estimated Variance	% variance
Among aureoles	2	9.858	4.929	0.087	2%
Within aureoles	46	167.183	3.634	3.634	98%
Total	48	177.041		3.721	100%

The results once again show that majority of the genetic variation is explained within aureoles (98%) compared to among aureoles (2%). The overall Φ_{PT} in Mali was 0.023 and slightly statistically significant P (rand >= data) of 0.01 and also as given by the frequency distribution or the randomised Φ_{PT} value versus the observed (**Figure 6.2**). However, the between aureoles comparisons did not yield statistically significant levels of genetic differentiation (**Table 6.4**).

Table 6.4 Between aureoles comparison of *P. falciparum* genotypes with the corresponding observed and permuted Φ_{PT} values shown.

Region comparison	Distance (km)	Observed Φ_{PT}	P (rand >= data)
Aureole 1 – Aureole 2	0.2	0.08	0.279
Aureole 1 – Aureole 3	0.2	0.044	0.084
Aureole 2 – Aureole 3	0.2	0.03	0.088

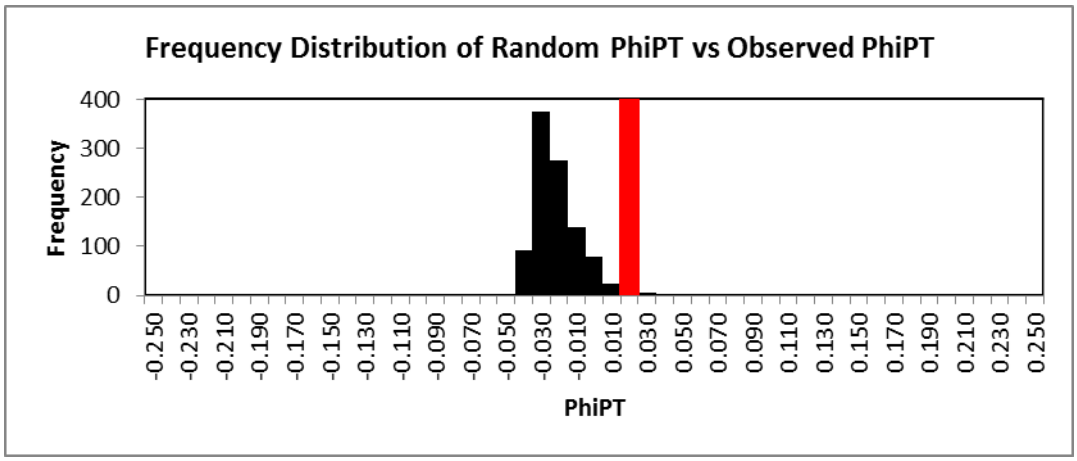


Figure 6.2 Frequency distribution of random Φ_{PT} (depicted in black) vs observed (in red) following randomisation as a test for statistical significance for the clonal genotypes obtained from Mali.

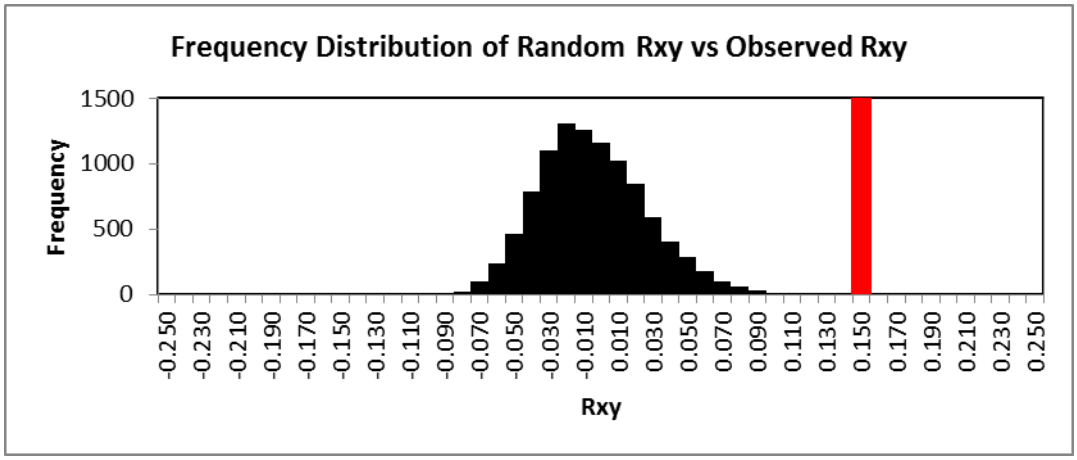
6.3.2. Mantel test

6.3.2.1. Overall data Mantel test

The Mantel test analysis of the overall data set yielded a correlation coefficient (R_{xy}) = 0.153, $P < 0.05$ for the actual geographic distances and 0.135, $P<0.05$ with the log transformed (**Figure 6.3 a&b)**. A slightly significant relationship by regression analysis (R^2 = 0.0234) was observed between the genetic and actual geographic distances of the parasite isolates. The analysis performed with the log transformed geographic distances as is recommended by Peakall and Smouse, (2006; 2012) showed an even less statistical significance by regression analysis (R^2 = 0.0183). Based on these observations, the H_0 was rejected but not strongly indicative of a

slightly significant relationship between the geographic distances and the genetic distances in the malaria parasite populations.

(a)



(**b**)

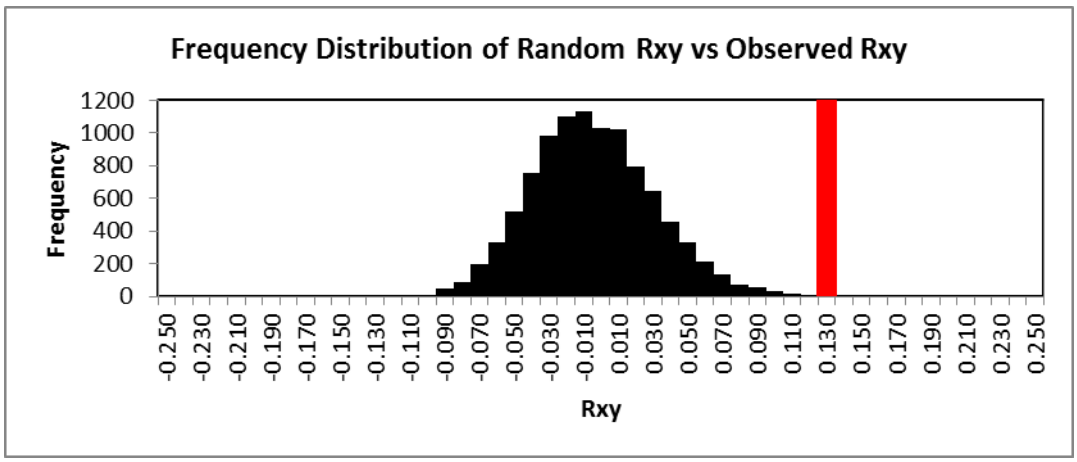
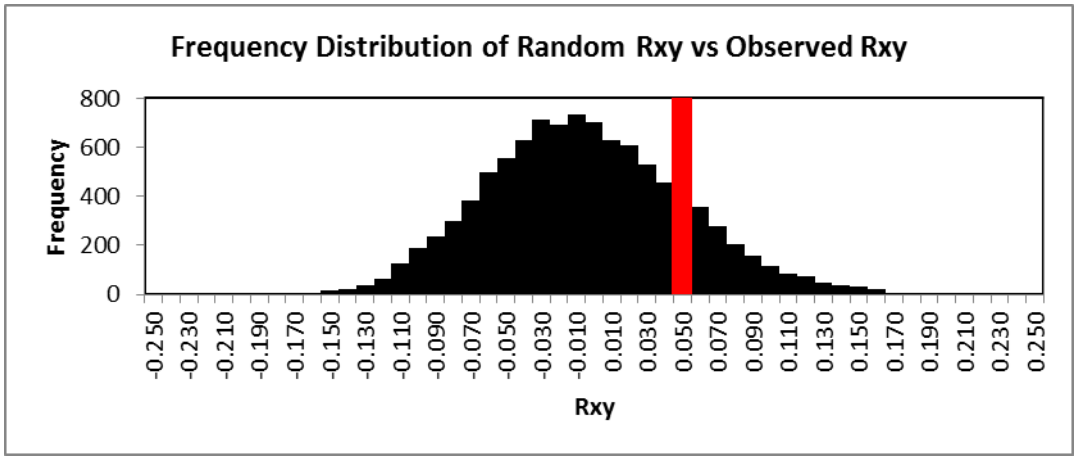


Figure 6.3 Frequency distribution of the correlation coefficient (Rxy) (depicted in black) vs observed (in red) following randomisation as a test for statistical significance for the clonal genotypes obtained over all the countries. The analyses performed using the actual geographic distances are given in (a) and the log transformed geographic distances in (b).

6.3.2.2. Mali study site Mantel test

The clonal genotypes obtained from Mali were also separately analysed for isolation by distance as the geographic co-ordinates were known. The results are shown in **Figure 6.4a & b.** No statistically significant relationship was observed between the genetic distance and the geographic distances with either the actual geographic distance analysis R^2 value = 0.0028 or with the log transformed R^2 = 0.0028. The obtained R_{xy} values = 0.053 for both the actual and log transformed geographic distances were also not statistically significant ($P < 0.05$) as shown in **Figure 6.4 a&b**. The results supported the null hypothesis (H_0) and it was therefore accepted.

(a)



(b)

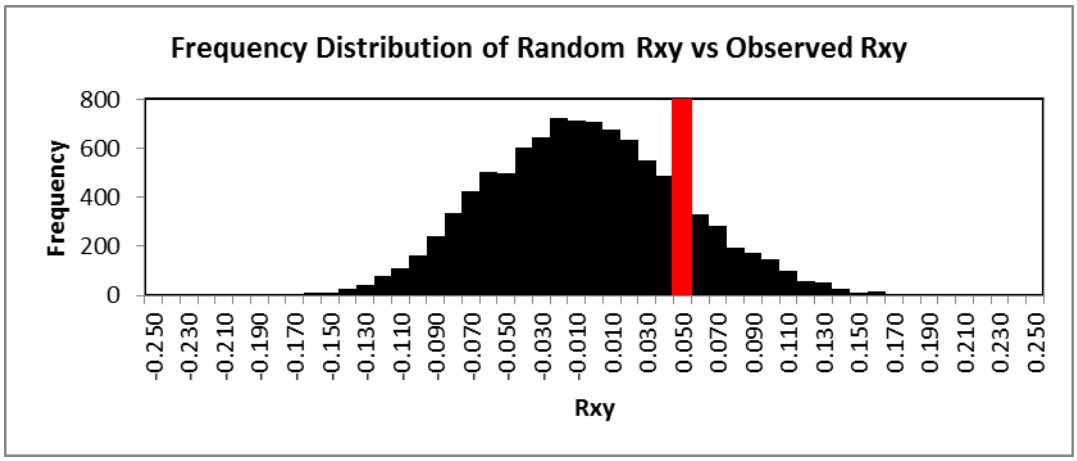


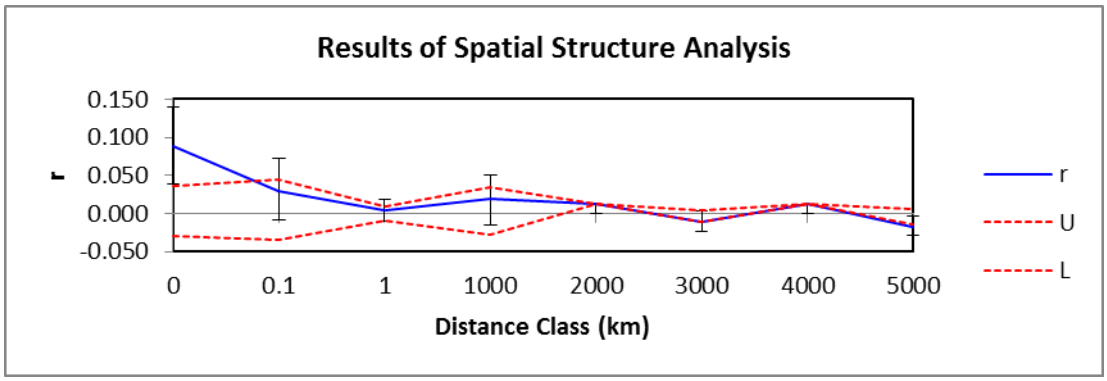
Figure 6.4 Frequency distribution of the correlation coefficient (Rxy) (depicted in black) vs observed (in red) following randomisation as a test for statistical significance for the clonal genotypes obtained from Mali. The analyses performed using the actual geographic distances are given in (a) and the log transformed geographic distances in (b).

6.3.3. Spatial autocorrelation analysis

6.3.3.1. Spatial autocorrelation analysis of the overall data set

The correlogram (**Figure 6.5**) obtained from the overall analysis of malaria parasite populations indicates that significant levels of spatial autocorrelation were detected within the first distance class (geographic distance between 0-0.1km). This is supported by the observed r value which exceeds the upper limit (upper red dashed line) and the mean bootstrap which exceeds 0 within this distance class. Thus, within geographic distance 0-0.1km, there seems to be an indication that parasites are

not randomly mating but are genetically structured. This might have been due to the sampling two genotypes from the same individuals that may be highly related genetically. This was investigated later and depicted in **Figure 6.7**.

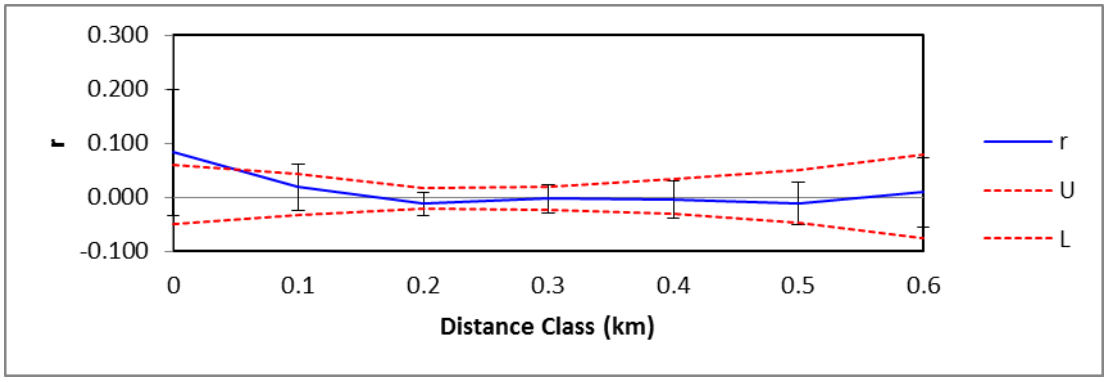


n	179	125	994	197	0	864	0	1127
Distance Class (km)	**0 - 0.1**	**0.1 - 1**	**1 - 1000**	**1000 - 2000**	**2000- 3000**	**3000 - 4000**	**4000 - 5000**	**5000 - 6000**
r	0.088	0.030	0.005	0.020	0.012	-0.011	0.012	-0.017
U	0.035	0.037	0.009	0.029	0.012	0.005	0.012	0.006
L	-0.029	-0.034	-0.007	-0.029	0.012	-0.010	0.012	-0.012
P(r-rand >= r-data)	0.001	0.049	0.121	0.084	1.000	0.981	1.000	0.994
P(r-rand <= r-data)	1.000	0.952	0.880	0.917	1.000	0.020	1.000	0.007
Mean Bootstrap r	0.087	0.030	0.005	0.019	0.000	-0.010	0.000	-0.017
Ur error	0.048	0.042	0.013	0.034	-0.012	0.015	-0.012	0.013
Lr error	0.050	0.041	0.013	0.036	0.012	0.013	0.012	0.013

Figure 6.5 Spatial autocorrelation analysis from the overall data. The correlogram depicts distance classes set in km as: 0-0.1, 0.1-1, 1-1000, 1000-2000, 2000-3000, 3000-4000, 4000-5,000 & 5000-6000. Although the distance classes are set up to 6000km (a total of 8 distance classes), the autocorrelation is performed for the 7 distance classes and any class beyond that is ignored. It is therefore pertinent to set the distance classes exceeding the maximum distance of separation of the isolates to ensure the whole data set is included in the analysis. The blue line represents the spatial autocorrelation coefficients computed within each distance class. The red dashed lines indicate the upper (U) 25th and lower (L) 975th confidence limits bound by the 95% confidence interval about the null hypothesis of no spatial structure for the combined data set as determined by permutation. When the blue line (r) exceeds the upper and lower confidence limits (red dashed lines), significant genetic structure is inferred. 1000 bootstrap trials give a mean bootstrap r bound by the 95% confidence limit. The upper (U) 25th and lower (L) 975th confidence limits are given by the error bars. When the mean bootstrap r is close to 0, the null hypothesis of no spatial structure is accepted. The table below this figure show the values obtained by both randomisation and bootstrapping analyses as tests for statistical significance.

6.3.3.2. Spatial autocorrelation analysis in Mali

Finally, a spatial autocorrelation analysis was performed for the isolates obtained from Kolle region, Mali (**Figure 6.6**) in an attempt to identify spatial genetic structure within short geographic distances.



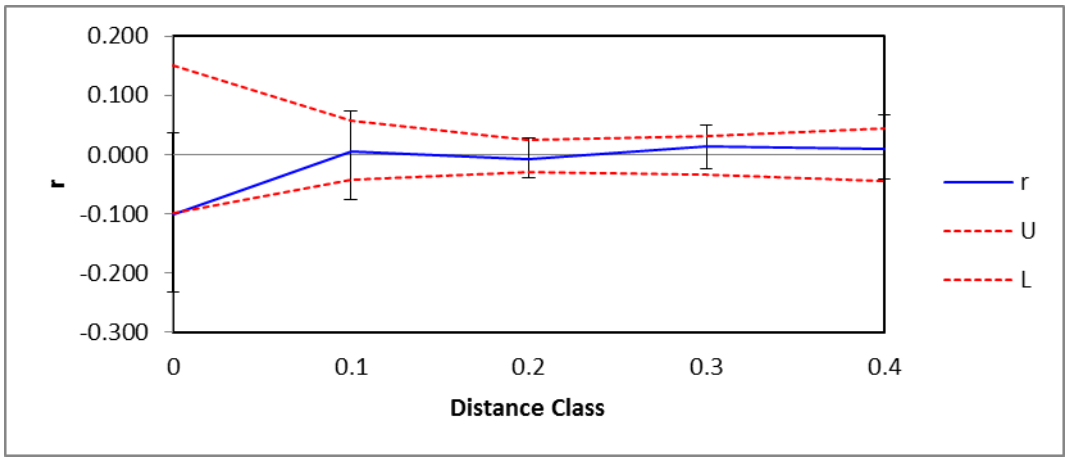
n	57	125	378	322	170	83	33
Distance Class (km)	**0 – 0.1**	**0.1 – 0.2**	**0.2 – 0.3**	**0.3 – 0.4**	**0.4 – 0.5**	**0.5 – 0.6**	**0.6 – 0.7**
r	0.083	0.019	-0.013	-0.002	-0.004	-0.011	0.009
U	0.067	0.043	0.018	0.021	0.032	0.047	0.075
L	-0.049	-0.035	-0.021	-0.026	-0.036	-0.050	-0.079
P(r-rand >= r-data)	0.012	0.159	0.906	0.572	0.626	0.694	0.433
P(r-rand <= r-data)	0.989	0.842	0.095	0.429	0.375	0.307	0.568
Mean Bootstrap r	0.084	0.018	-0.012	-0.002	-0.004	-0.011	0.009
Ur error	0.120	0.045	0.022	0.026	0.036	0.038	0.065
Lr error	0.113	0.043	0.022	0.026	0.037	0.040	0.060

Figure 6.6 A correlogram of spatial autocorrelation analysis for Kolle region, Mali data. This illustration shows distance classes set in km as: 0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5, 0.5-0.6 & 0.7-0.8. Once again, the spatial autocorrelation is performed up to the distance class (0.5-0.6) and the last one is ignored. The blue line represents the spatial autocorrelation coefficients computed within each distance class. The red dashed lines indicate the upper (U) 25th and lower (L) 975th confidence limits bound by the 95% confidence interval about the null hypothesis of no spatial structure for the combined data set as determined by permutation. When the blue line (r) exceeds the upper and lower confidence limits (red dashed lines), significant genetic structure is inferred. 1000 bootstrap trials give a mean bootstrap r bound by the 95% confidence limit. The upper (U) 25th and lower (L) 975th confidence limits are given by the error bars. When the mean bootstrap r is close to 0, the null hypothesis of no spatial structure is accepted. The table below shows the r

values obtained by both randomisation and bootstrapping analyses as tests for statistical significance.

The combined SNP and microsatellites data set of *P. falciparum* clonal genotypes obtained from Mali showed statistically significant spatial structuring within the 1st class (0-0.1km) (**Figure 6.6**). Within this class, the H_0 was rejected as there was clear indication of genetic structuring as revealed by the r value (blue line) exceeding the upper limit (upper red dashed line) in addition, the mean bootstrap r was above 0.

However, the analysis performed in **Figure 6.6** included within individual clonal genotypes which are usually highly related genetically compared to infections obtained between individuals. To ascertain whether these results reflect actual genetic structuring of malaria parasites within the observed geographic distance, analysis comprising the major clones only in Mali was carried out. The results are shown in **Figure 6.7**. These results are a clear indication that the observed genetic relatedness in **Figure 6.6** was due to the existence of within individuals parasite infections in the analysis since analysis with the major clones only did not give any evidence of genetic structuring within short geographic distances.



n	16	64	176	146	82
Distance Class (End Point)	**0-0.1**	**0.1-0.2**	**0.2-0.3**	**0.3-0.4**	**0.4-0.5**
r	-0.101	0.005	-0.007	0.014	0.010
U	0.150	0.057	0.024	0.032	0.045
L	-0.099	-0.044	-0.030	-0.033	-0.044
P(r-rand >= r-data)	0.977	0.386	0.733	0.198	0.320
P(r-rand <= r-data)	0.024	0.615	0.268	0.803	0.681
Mean Bootstrap r	-0.089	-0.001	-0.007	0.014	0.012
Ur error	0.136	0.069	0.034	0.035	0.058
Lr error	0.133	0.081	0.032	0.038	0.052

Figure 6.7 A correlogram of spatial autocorrelation analysis for Kolle region, Mali comprising data from the major clones only. This illustration shows distance classes set in km as: 0-0.1, 0.1-0.2, 0.2-0.3, 0.3-0.4, 0.4-0.5 & 0.5-0.6. Once again, the spatial autocorrelation is performed up to the distance class (0.4-0.5) and the last one is ignored. The blue line represents the spatial autocorrelation coefficients computed within each distance class. The red dashed lines indicate the upper (U) 25th and lower (L) 975th confidence limits bound by the 95% confidence interval about the null hypothesis of no spatial structure for the combined data set as determined by permutation. When the blue line (r) exceeds the upper and lower confidence limits (red dashed lines), significant genetic structure is inferred. 1000 bootstrap trials give a mean bootstrap r bound by the 95% confidence limit. The upper (U) 25th and lower (L) 975th confidence limits are given by the error bars. When the mean bootstrap r is close to 0, the null hypothesis of no spatial structure is accepted. The table below shows the r values obtained by both randomisation and bootstrapping analyses as tests for statistical significance.

6.4. Discussion

In this concluding chapter of the thesis, genetic distance based analyses were performed. With reference to the concept forming the basis of this project outlined in **Chapter 1, Section 1.6**, the analyses described in this chapter attempted to address the distribution of malaria parasites in space. It is already known that malaria parasite transmission is heterogeneous. Elucidation of parasite transmission patterns in space would aid in determining discrete clusters of parasite populations in

different geographic regions. These clusters can then be identified and targeted for stratified malaria elimination programs.

Spatial distribution of malaria parasites has been investigated prior to establishment of malaria elimination programs previously. Geographical reconnaissance using GPS to determine high malaria risk areas (hotspots) was used to develop malaria risk maps in a baseline survey conducted at the Solomon islands and Vanuatu islands (Reid *et al*., 2010; Kelly *et al*., 2010). Other studies have attempted to establish clusters of malaria transmission using sero-conversion rates of antibodies (Bousema *et al*., 2010; Bejon *et al*., 2011).

The use of neutral genetic markers to detect clusters of closely mating parasites has been attempted in few studies. These have been conducted using microsatellites in Brazil (Machado *et al*., 2004), Malaysia (Anthony *et al*., 2011); and recently in West Africa (Mobegi *et al*., 2012). SNP assays have also been developed and tested for *P. falciparum* population genetic studies (Campino *et al*., 2011; Manske *et al*., 2012). So far, detection of the genetic relatedness of malaria parasites to trace parasite transmission patterns in space has not been attempted. It is on this basis that this study was developed, to: 1. Identify suitable genetic markers that can isolate malaria parasite clones present in mixed clone infections, 2. Use the identified clonal genotypes to unravel the spatial patterns of *P. falciparum* within short geographic settings (<1km) and across different African countries. The information obtained in this study can subsequently be used in future studies to trace transmission of these parasites from one individual to another. This would aid in the identification of

discrete clusters of malaria transmission in space that can be targeted in malaria intervention programs.

6.4.1. AMOVA analysis

The results obtained in this study reveal that across the 3 African countries namely, Cameroon, Kenya and Mali majority of the genetic variation is explained among parasites within populations rather than between the 3 countries (**Table 6.1**). Similar findings were observed in a study comparing *P. falciparum* populations between a lowland and highland region using microsatellites estimation of allele frequencies in Western Kenya. This study observed that between the two regions, 92.2% of the variation was explained within *P. falciparum* populations and 6.9% among populations (Zhong *et al*., 2007).

The results in **Table 6.1** give further evidence to the pre-existing knowledge of observed significant parasite genetic diversity levels in Africa's high malaria transmission areas (Anderson *et al*., 2000). The overall Φ_{PT} value observed was low and in concordance with F_{ST} values observed between parasite populations across Africa (Anderson *et al*., 2000; Mobegi *et al*., 2012). The overall observed Φ_{PT} (0.049) is also comparable to the overall θ value obtained using microsatellites only (0.043).

The low Φ_{PT} value obtained in this study, in addition to the low F_{ST} values obtained in other studies, give an impression of randomly mating parasite populations across distances >1000km. This is indeed not possible considering that the greatest distance

of dispersal for the mosquito is approximately 1km (Trape *et al*., 1992; Takken *et al*., 1998) (not considering mosquitos implicated for spread of 'airport malaria'). Human movement between sites has also been implicated in facilitating mixing of malaria parasites leading to low observed F_{ST} values (Zhong *et al*., 2007). The most plausible reason for the low observed values of genetic differentiation is the presence of highly genetically diverse parasite clones within multiply infected human hosts such that the effective population size (number of parasites that are mating and contributing to the gene pool) is increased. Thus, low levels of genetic differentiation are maintained because with such large effective population sizes the effects of genetic drift are minimal. In addition, migration of infected individuals even to within short geographic distances constributes substantially to the existing parasite gene pool as a result of the high rate of recombination between genetically distinct clones. Although there could be existing clusters of closely mating parasites they might be overlapping within short geographic distances diminishing the genetic structuring of these parasites.

6.4.2. Mantel test

A low but statistically significant positive correlation of genetic and geographic distances was observed in the *P. falciparum* clonal genotypes from the 3 African countries (**Figure 6.3**). This is in conformity with the isolation by distance theory stating that the genetic relatedness of malaria parasites reduces with increasing geographic distance. This is what is typically expected in nature and has been widely studied in plants (reviewed by Vekemans & Hardy, 2004). This was however not

observed in Mali giving the impression of a randomly mating parasite population within the study site (**Figure 6.4**).

Isolation by distance phenomenon has been observed in *P. falciparum* in a region experiencing reductions of malaria incidences in Borneo, Malaysia (Anthony *et al*., 2005). However, in West Africa no correlation was observed between geographic distances and genetic distances (Mobegi *et al*., 2012). The failure to detect isolation by distance in Mali could be indicative of no genetic structuring or existence of discrete clusters of closely mating parasites that could not be elucidated with the set of genetic markers used in this study.

6.4.3. The spatial autocorrelation test

This analysis was performed to examine at finer geographic spatial scales the distribution of malaria parasites. The spatial autocorrelation analysis performed over all the *P. falciparum* clonal genotypes (**Figure 6.5**) and in the Malian genotypes, revealed no spatial autocorrelation except within a region spanning 0-0.043km (**Figure 6.6**). This observation is not surprising due to the existence of closely related parasite genotypes obtained within a single isolate. A recent study by Nkhoma *et al*., 2012 observed that comparison between *P. falciparum* haplotypes recovered within infections were closely related genetically than those between infections. Since genotypes obtained within and between infections were included in this analysis, within infection clonal genotypes obtained from a single isolate are more genetically related compared to between infection clonal genotypes. This was

further evidenced in the analysis using major clones only in Mali (**Figure 6.7**) whereby no genetic structuring was observed.

Although in **Chapter 5** significant levels of spatial structuring were observed in Kolle, Mali within geographic distances ~0.4km, the results obtained in this chapter indicate that there is no population structure in this region meaning it comprises randomly mating parasites. Perhaps there is considerable human or mosquito migration facilitating mixing of malaria parasites.

It is noteworthy that the findings obtained in the Kolle region, Mali only pertain to this study site and is not a reflection of the overall *P. falciparum* spatial patterns across all African sites. The dynamics of malaria transmission are varied and specific to each study site. Indeed this study recognizes the lack of actual geographic co-ordinates for the isolates collected from Rusinga Island, Kenya or in Cameroon. This information would have provided alternative spectra in defining patterns of malaria parasites' distribution within small spatial scales. Even so, performing a combinatorial analysis using both genetic markers has given a snapshot view of the spatial distribution of malaria parasites within a region separated by distances (<1km), and across Africa.

6.4.4. Conclusion

This chapter culminates the analyses planned for this thesis. The genetic distance based analyses outlined in this chapter have revealed how malaria parasites are

dispersed in space. No significant spatial autocorrelation has been observed within and between the studied sites in Africa

7. FINAL DISCUSSION

7.1. Synopsis of thesis

This thesis entailed a feasibility study to investigate the transmission patterns of *P. falciparum* malaria parasites in short geographic distance (~1km). The steps involved in achieving this are outlined below.

1. This study intended to use genotyping methods for the identification of infecting malaria parasites present in human blood. The first step was therefore, to find polymorphic genetic markers in this case, SNPs. The process involved in the identification and ascertainment of these SNPs is covered in **Chapters 3 & 4**.
2. The next step was to establish a method for SNP detection that enabled the identification of parasite genotypes. Considering that malaria parasites exist as mixed clone infections, a technique facilitating the differentiation of clonal genotypes from these infections was required. Pyrosequencing^{TM}, a technique that identifies SNPs and quantifies them assigning proportions relative to their amounts e.g. in a mixed clone infection, was used. Parasite clonal genotypes could then be determined on the basis that SNPs occurring in the same proportion comprised genotypes representing a single clone. The use of the identified SNPs and technique validation is discussed in **Chapters 3 & 4**.
3. Having identified the SNPs and validated the technique, analysis of *P. falciparum* field isolates was carried out. Malaria positive blood samples obtained from Kenya, Cameroon and Mali were analysed. Initially, sample

collections were done in Kenya, the major criterion in this site being that they should be obtained from inhabitants of Rusinga Island. However, in the course of the study we realised that parasite isolates obtained from within short geographic distances of each other (~1km) and furnished with geographic co-ordinates would be more informative for the purposes of this thesis. Thus, further collections were carried out in Mali within a region of <1km and geographic coordinates were available. In addition, through collaboration between the University of Edinburgh and the University of Cameroon, further collections were carried out in Cameroon to investigate the feasibility of using Illumina - Solexa sequencing for parasite clonal genotypes identification. Use of this technique was later rejected due to financial constraints and the amount of post-sequencing *in silico* work required. This aspect was however not covered in this thesis. Nevertheless, the Cameroon samples were also analysed in this thesis even though information on their geographic coordinates was missing.

In addition to the SNPs, microsatellites were also used to increase the genotypic information obtained using SNPs only, which facilitated the identification of more clonal genotypes. However, since both sets of genetic markers have different mutation rates, they were analysed separately in the model based analyses outlined in **Chapter 5**.

Using the identified parasite clonal genotypes, informative tests were performed to reveal the usefulness of the SNPs and microsatellites in the detection of genetic variation in field isolates. Also, this study intended to reveal the gene flow patterns of these parasites in different geographic

settings. For this purpose, neutral genetic markers are preferred over use of markers that could be under selection. This is because selection alters patterns of allelic diversity obscuring the effects of gene flow patterns in the distribution of allelic diversity in space. The microsatellites were tested for presence of LD which, if present, is indicative of the non-random association of alleles due to selection. They were found not to be in LD (**see Chapter 5**). Evidence of population structure in the regions studied was investigated. This was performed to reveal patterns of *P. falciparum* transmission whereby partitioning of parasites in a population into smaller subpopulations with limited gene flow between them would reveal that majority of the genetic diversity is explained between the subpopulations as opposed to within them. In addition, the partitioning of individual parasites into small isolated subpopulations leads to their being independently acted upon by genetic drift altering their allele frequencies and further increasing the genetic divergence between them.

At a continental level, statistically significant levels of population structuring were observed by comparing clonal genotypes obtained from Kenya, Mali and Cameroon. This is expected as at this level, parasites are separated by large geographic distances with limited gene flow between them. Further partitioning of the clonal genotypes obtained from the Mali study site whose: (i) geographic coordinates were known and (ii) isolates were obtained from within a distance of <1km was performed and genetic differentiation determined. Both sets of markers of markers seemed to indicate theat there

was evidence of genetic differentiation within a distance of 400m (**Chapter 5**).

4. Finally, a spatial autocorrelation analysis was performed in an attempt to reveal distinct clusters of closely mating parasites in space. For these analyses, the SNPs and microsatellites loci were combined. At a continental level, a slightly significant positive correlation between genetic and geographic distances was observed. However, a similar test in Mali (region <1km) revealed no correlation. Further to this, the spatial autocorrelation test revealed no spatial structuring in Mali (**Chapter 6**).

7.2. Key findings and their implications in malaria management

This study was an attempt to detect levels of genetic relatedness in malaria parasite isolates occurring within short geographic distances. The findings in this study indicate that no closely mating parasites within tightly clustered regions were observed. So, what does this mean?

Currently, there is a renewed commitment globally to reduce or even eradicate malaria incidences. To achieve this, we need to borrow lessons from the management and successful eradication of other infectious diseases.

To date, the only disease to be successfully eliminated is small pox although other concerted efforts have led to reductions but not eradication of polio and measles. Firstly, these programs are evidence-based. For example in the management of measles; genotyping identified the genetic relatedness of the virus in different geographic settings allowing the monitoring of vaccine failures, disease resurgences

and its transmission patterns e.g. the occurrence of imported cases (Mulders *et al*., 2001; Tischer *et al*., 2004) The role of research in the active surveillance of disease management programs is therefore relevant. Malaria intervention studies currently appreciate this which is one of the reasons why the study described in this thesis was formulated. This project has helped me appreciate the importance of evidence-based malaria intervention strategies. It has also served as a good foundation as I take up my post-doctoral fellowship position on a project involving malaria elimination in the Lake Victoria Islands, Kenya. My roles include determining the gene flow patterns of the malaria patterns due to migration of either people or mosquitos within and between the islands. In addition, the *P. falciparum* multiplicity of infection will be determined before and after intervention to monitor the effectiveness of the control strategy. Successful interventions leading to malaria elimination in these high malaria transmission islands will hopefully inform planned malaria studies in the mainland.

Secondly, successful eradication of small pox and type 2 polio was possible following the development of vaccines for immunization (Andre *et al.*, 2008). Current malaria research is focussing on the development of transmission blocking vaccines that would curtail transmission of the infective stages of malaria parasites by the mosquito. Unlike an asexual stage vaccine which protects an individual from the development of clinical disease, the transmission blocking vaccine has the added advantage of protecting the community by reducing parasite progression which is the basis of vaccine development. The focality of malaria transmission further enhances its efficacy because the immunisation of one susceptible individual not only effectively reduces the efficiency of parasite transmission within his/her house but

also in the neighbouring houses a phenomenon referred to as 'herd immunity.' This is the "reduction of infection or disease in the unimmunised segment as a result of immunising a segment of the population" (John & Samuel, 2000). This in turn leads to the reduction of R_0 within a household and subsequently in the community.

In this study however, the existence of tightly clustered malaria parasites was not observed. This means that either the paucity of the genetic markers used led to failure in the detection of malaria parasites' divergence occurring within the studied regions or perhaps the level of genetic diversity in these parasites is so high (due to their high effective population size) that it is difficult to observe population structure. Based on the findings of other studies carried out in Africa e.g. Anderson *et al*., 2000, the latter explanation seems highly plausible. It is also possible that the current malaria control efforts in Africa have still not managed to reduce malaria endemicity to levels enabling the fragmentation of parasites to small closely mating subpopulations.

7.3. Lessons learnt

The overall objective of this study i.e. the exploration of malaria parasites' distribution within small geographic settings was achieved. However, important issues that arose in the course of the study are outlined below.

1. Proper project design – The sampling strategy whereby samples were initially collected from Kenya with no geographic coordinates which necessitated further collections in a different study site highlights the need for ample planning in the design of the project. This would have saved time and costs.

In addition, the Kenyan samples were not informative in the intended study objective.

2. Few genetic markers – Although the identified SNPs for use in this study were observed to be informative, higher genome coverage of *P. falciparum* genetic markers is required. I appreciate that other studies have identified suitable SNPs for population genetics studies using high density genome-wide SNP coverage e.g. Manske *et al*., 2012. To publish this work, only microsatellites data will be considered.
3. Although ample information was obtained using both SNPs and microsatellites for the identification of clonal genotypes for future studies it is better to use one set of genetic markers not both of them combined especially to make inferences in population genetic studies. As already stated in **Chapter 5**, the fact that they have different mutation rates necessitates their separate analysis.

   It is noteworthy that a study can be designed to accommodate both sets of genetic markers. For example Nkhoma *et al*., (2012) used microsatellites to identify mixed clone infections in field isolates. From this, they isolated single clones by a semi-cloning step and subsequently used SNPs to identify haplotypes for the determination of parasite genetic relatedness.

7.4. Future prospects of study

This feasibility study has ascertained that this work can be done successfully with proper design, well thought out sampling strategy and with adequate number of

genetic markers. However, it is noteworthy that with the current malaria transmission patterns in Africa, it is not possible to detect patterns of genetic differentiation yet. With the ongoing persistent malaria control efforts, regional fragmentation into patches of varying transmission intensities is expected to emerge. It is at this point that a study of this nature will be highly applicable.

This section therefore highlights future considerations in the design of a project entailing tracing of malaria parasites in space as outlined in the project concept (**Chapter 1, Section 1.6**).

1. Study site

   A suitable study site would include human habitation and proximity to a mosquito breeding site within the estimated average distance of travel for a mosquito to and from a breeding site (~1km). The closeness of the houses also needs to be factored in and also the number of individuals per house. This is to reveal patterns of malaria parasites transmission from person to person (e.g. within a single household) or between houses in close proximity of each other.

2. Sampling strategy

   Once an appropriate study site has been established, the households have been properly mapped and the geographic coordinates are known, it is important to consider the number of times sampling will be done to ensure adequate representation of infecting parasites for population genetic inferences.

   To begin with, the transmission intensity of an area needs to be factored in. For example in a high malaria transmission area, it can be assumed that

individuals are continuously bitten by the mosquito and inoculated with parasites. On the other hand in a low malaria transmission region, the rate of parasite inoculations is scanty.

Secondly, since sampling precedures may miss out sequestered parasite clones, repeated sampling is relevant. The issue of how often this needs to be done arises. Considering the *P. falciparum* parasite lifecycle (see **Chapter 1, Figure 1.1**) and assuming that malaria transmission is continuous as in a high malaria intensity region see outlined time points (**Figure 7.1**).

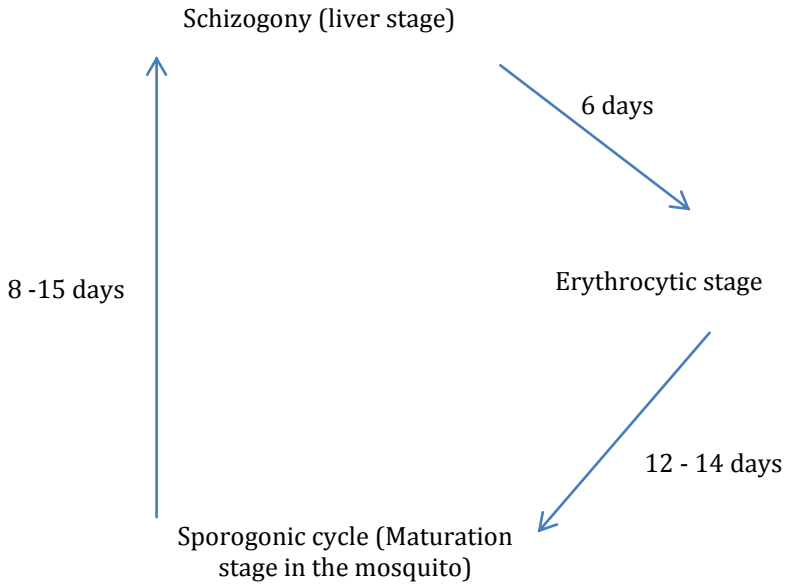


Figure 7.1 Time to maturation of malaria parasites in the human host and in the mosquito. In the human host, it takes ~6 days from the time malaria parasites are inoculated to the human host to their appearance in peripheral blood (schizogony to erythrocytic stage) at which point they can be collected for analysis. For gametocyte maturation, it takes several 48 hour proliferative asexual cycles for the peak parasitaemias in the human host to be achieved triggering a stress response that initiates gametocyte formation. This has been factored in the gametocyte maturation period to their ingestion by the mosquito but it may take longer than this (12-14 days). It then takes another 8-15 days in the mosquito for the infective sporozoites to mature and be reintroduced to the human host (sporogonic cycle). It then takes another ~6 days before these parasites can be detected in peripheral blood for sampling.

According to **Figure 7.1**, it takes ~32 days for a complete cycle from the first infective bite to the next reinfection. This is assuming that the mosquito lives long enough to infect and re-infect a person. This is important because, if sampling is done at time 0 and clonal genotypes obtained, ~32 days later novel genotypes can be produced following genetic recombination in the mosquito. Thus parasite genotypes that may have been missed out due to sequestration can possibly be detected as products of recombination. With repeated sampling, a clearer picture will emerge between parental genotypes and the progeny. The question of the number of times this needs to be done will be covered in the next section.

3. Genetic markers

Although both SNPs and microsatellites were used in this thesis for the identification of parasite clonal genotypes, SNPs were the choice genetic markers for analysis as outlined in **Chapter 3** and will be discussed in this section. In order to trace how malaria parasites are being transmitted in space, it is important to use high density SNPs per chromosome for haplotype (blocks of adjacent nucleotides that are inherited together) detection. Using these haplotypes, closely related parasites can therefore be identified. However, these blocks are usually broken down by recombination since *P. falciparum* has a high recombination rate i.e. ~15-20kb/cM per generation such that haplotype blocks have a high rate of being broken down. This is further complicated by the high rates of multiple infections in regions of high malaria transmission intensity leading to increased numbers of novel clone genotypes. How dense then should the SNPs coverage be per chromosome?

A study by Mu *et al*., (2005) revealed that different rates of recombination occurred on the *P. falciparum* chromosome 3 (size ~1MB) with the highest rates observed in

the sub-telomeric regions. In this study, a haplotype block of 11.2kb was obtained in African *P. falciparum* populations. It is therefore recommended that high SNP coverage from short and longer chromosomes be used. This is because in longer chromosomes, the rates of recombination relative to the chromosomal sizes have been observed to be lower than in shorter chromosomes (Mu et al, 2007) increasing the chances of haplotype detection.

4. Detection method

The Pyrosequencing^{TM} technique developed and used in this thesis is appropriate for this analysis more so because it allows the inference of relative proportions of genetically distinct clones per infection. However, the process of generating assays is long and time consuming including its limits of accuracy as outlined in **Chapter 4**. The advances on next generation sequencing offer alternative methods in the analysis of genome-wide SNPs. However, as already stated, when dealing with mixed clone infections the identification of genotypes is still a challenge. Although it is time consuming, a semi-cloning step such as that used by Nkhoma *et al*., (2012) can be incorporated in the analysis to ensure that only a single parasite clone per isolate is sequenced. This eases the subsequent *in silico* post-sequencing identification of clonal genotypes.

5. Interpretation of results

Once the haplotypes have been identified, one of the challenges in the interpretation of results is defining alleles that are identical by state (IBS) or identical by descent (IBD). To understand this concept, see **Figure 7.2** below.

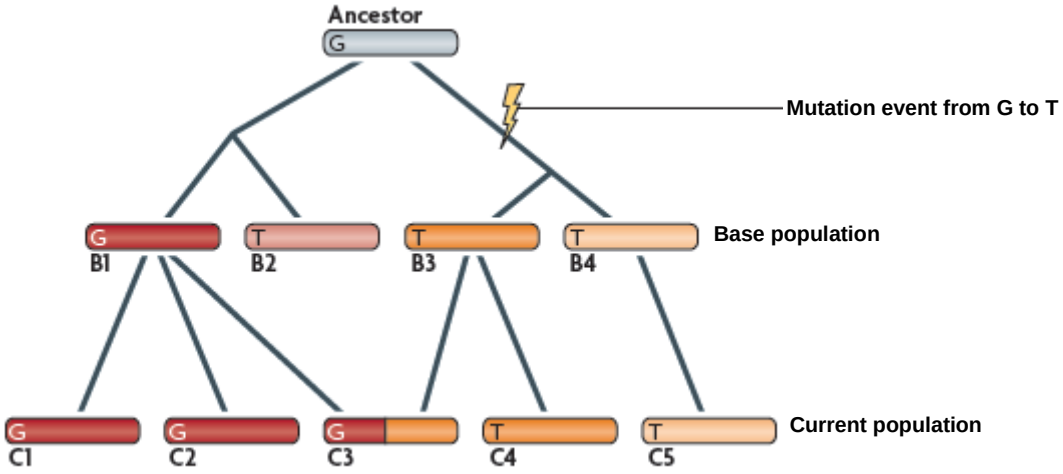


Figure 7.2 Depiction of IBD and IBS. The current population can be traced back to its original ancestor by coalescence. The ancestral chromosome has a polymorphic G/T region i.e. the 'G' allele mutates to a 'T'. The base population when performing coalescence simulations comprises individuals assumed to be unrelated hence the multi-coloured chromosomes. In the current population the G alleles in the C1, C2 and C3 chromosomes are IBD as they are inherited from the same ancestor (B1). On the other hand, the T allele carried by C4 and C5 are IBS, but not IBD, as they descend from different alleles in the base population (B3 and B4, respectively). (Figure adapted from Powell *et al*., 2010).

Figure 7.2 reveals the complexities involved in the interpretation of data involved in a pedigree analysis (involving mating patterns of closely related individuals) which is appropriate for the project concept outlined in **Chapter 1, Section 1.6**. It is expected that mating among relatives (inbreeding) is concordant with high observed GR values. However, in this analysis, the current population (the first sampled population) is usually assumed to constitute the parental clones and probabilistic inferences made on their ancestral population (base population in **Figure 7.2**). It is at this point that erroneous interpretation can be made if IBS alleles are not IBD.

This problem can be circumvented by using highly dense continuous genetic markers per chromosome for accurate detection of haplotypes that are IBD (Browning, 2008).

Ths is because highly dense SNPs per chromosome have a high chance of being IBD due to linkage obscuring the effects of IBS alleles. The use of high density SNPs coupled with sampling at more than one time point would ensure accurate detection of haplotypes from which inferences on the local mating patterns of parasites can be made.

In conclusion, the findings and the lessons learnt in this study offer tremendous prospects to the tracing of malaria parasites within short geographic distances revealing the transmission patterns of these parasites from person to person.

8. REFERENCES

Adjei O, May J (2008). Spatial variation of malaria incidence in young children from a geographically homogeneous area with high endemicity. *The Journal of Infectious Diseases* **19**7(1): 85–93

Albrechtsen A Nielsen FC, Nielsen R (2010). Ascertainment biases in SNP chips affect measures of population divergence. *Molecular Biology and Evolution* **27**(11): 2534–47

Alonso PL, Brown G, Arevalo-Herrera M, Binka F, Chitnis C, Collins F, Doumbo OK, Greenwood B, Hall FM, Levine MM, Mendis K, Newman DR, Plowe CV, Rodriguez MH, Sinden R, Slutsker L, Tanner M (2011). A research agenda to underpin malaria eradication. *PLoS Medicine* **8**(1):1-8

Anderson TJ, Su XZ, Bockarie M, Lagog M, Day, KP (1999). Twelve microsatellite markers for characterization of *Plasmodium falciparum* from finger-prick blood samples. *Parasitology* **119** (2): 113–25

Anderson TJ, Haubold B, Williams JT, Estrada-Franco GJ, Richardson L, Mollinedo R, Bockarie M, Mokili J, Mharakurwa S, French N, Whitworth J, Velez DI, Brockman AH, Nosten F, Ferreira MU, Day KP(2000). Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*. *Molecular Biology and Evolution* **17**(10): 1467–82

Anderson, TJC, Nair S, Sudimack D, Williams JT, Mayxay M, Newton PN, Guthmann JP, Smithuis FM, Tran TH, van den Broek IVF, White NJ, Nosten F (2005). Geographical distribution of selected and putatively neutral SNPs in Southeast Asian malaria parasites. *Molecular Biology and Evolution* **22**(12): 2362–74

Andre FE, Booy R, Bock H, Clemens J, Datta SK, John TJ, Lee BW, Lolekha S, Peltola H, Ruff TA, Santosham M, Schmitt HJ (2008). Vaccination greatly reduces disease, disability, death and inequity worldwide. *Bulletin of the World Health Organization* **86**(2): e1000405

Ansell J, Hamilton KA, Pinder M, Walraven GEL, Lindsay S (2002). Short-range attractiveness of pregnant women to *Anopheles gambiae* mosquitoes. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **96**(2):113–6

Anthony TG, Conway DJ, Cox-Singh J, Matusop A, Ratnam S, Shamsul S, Singh B. (2005). Fragmented population structure of *Plasmodium falciparum* in a region of declining endemicity. *The Journal of Infectious Diseases* **191**(9): 1558–64

Arez AP, do Rosário, VE (2008). The relevance of molecular markers in the analysis of malaria parasite populations. *Transboundary and Emerging Diseases* **55**(5-6): 226–32

Arnot DE, Roper C, Bayoumi R.AL (1993). Digital codes from hypervariable tandemly repeated DNA sequences in the *Plasmodium falciparum* circumsporozoite gene can genetically barcode isolates. *Molecular and Biochemical Parasitology* **61**(1): 15–24

Arnot DE (2002). The influence of the genetic complexity of *Plasmodium falciparum* infections on the epidemiology of malaria. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **96**(1): S1/131–S1/136

Auburn S, Campino S, Clark TG, Djimde AA, Zongo I, Pinches R, Manske M, Mangano V, Alcock D, Anastasi E, Maslen G, MacInnis B, Rockett K, Modiano D, Newbold CI, Doumbo OK, Ouédraogo JB, Kwiatkowski DP (2011). An Effective Method to Purify *Plasmodium falciparum* DNA Directly from Clinical Blood Samples for Whole Genome High-Throughput Sequencing. *PLoS ONE* **6**(7): e22213

Aujoulat F, Roger F, Bourdier A, Lotthé A, Lamy B, Marchandin H, Jumas-Bilak (2012). From Environment to Man: Genome Evolution and Adaptation of Human Opportunistic Bacterial Pathogens. *Genes* **3**: 191–232

Avise CJ (2004). Short tandem repeat loci (STRs) (microsatellites). In Molecular markers, natural history and evolution (2nd eds), pp 92-93. Sinauer Associates, Sunderland, Massachussets

Ayala FJ, Escalante AA, Lal AA, Rich SM (1998) in *Malaria:Parasite Biology, Pathogenesis, and Protection*, ed. Sherman, I. W. (American Society of Microbiology, Washington, DC), pp. 285–300.

Babiker HA, Ranford-Cartwright LC, Currie D, Charlwood JD, Billingsley P, Teuscher T, Walliker D (1994). Random mating in a natural population of the malaria parasite *Plasmodium falciparum*. *Parasitology* **109**(4): 413–21

Babiker HA, Walliker D (1997). Current views on the population structure of *Plasmodium falciparum*: Implications for control. *Parasitology Today* **13**(7): 262–7

Babiker HA, Abdel-Muhsin AM, Ranford-Cartwright LC, Satti G,Walliker D (1998). Characteristics of *Plasmodium falciparum* parasites that survive the lengthy dry season in eastern Sudan where malaria transmission is markedly seasonal. *The American Journal of Tropical Medicine and Hygiene* **59**(4): 582–90

Babiker HA, Ranford-Cartwright LC, Walliker D (1999). Genetic structure and dynamics of *Plasmodium falciparum* infections in the Kilombero region of Tanzania. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **93**: S11–S14

Baird JK, Jones TR, Danudirgo EW, Annis BA, Bangs MJ, Basri H, Purnomo Masbar S (1991). Age-dependent acquired protection against *Plasmodium falciparum* in people having two years exposure to hyperendemic malaria. *The American Journal of Tropical Medicine and Hygiene* **45**(1): 65–76

Balloux F, Lugon-Moulin N. (2002). The estimation of population differentiation with microsatellite markers. *Molecular Ecology* **11**(2): 155–65

Beier JC (1998). Malaria parasite development in mosquitoes. *Annual Review of Entomology* **43**: 519–43

Bejon P, Williams TN, Liljander A, Noor AM, Wambua, J, Ogada E, Olotu A, Osier F HA, Hay SI, Färnert A, Marsh, K (2010). Stable and unstable malaria hotspots in longitudinal cohort studies in Kenya. *PLoS medicine* 7(7): p.e1000304

Bejon P, Turner L, Lavstsen T, Cham G, Olotu A, Drakeley C J, Lievens M, Vekemans J, Savarese G, Lusingu J, Seidlein L, Bull P, Marsh K, Theander GT (2011). Serological Evidence of Discrete Spatial Clusters of *Plasmodium falciparum* Parasites. *PloS one* **6**(6): e21711. doi:10.1371/journal.pone.0021711

Bendixen M, Msangeni HA, Pedersen BV, Shayo D, Bødker R (2001). Diversity of *Plasmodium falciparum* populations and complexity of infections in relation to transmission intensity and host age: a study from the Usambara Mountains, Tanzania. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **95**(2): 143–8

Berendt AR, Ferguson DJP, Gardner J, Turner GT, Rowe A, McCormick C, Roberts D, Graig A, Pinches R, Elford BC, Newbold CI (1994). Molecular mechanisms of sequestration in malaria. *Parasitology* **108**:S19-S28

Bhattarai, A, Ali AS, Kachur SP, Mårtensson A, Abbas AK, Khatib R, Al-mafazy A-W, Ramsan M, Rotllant G, Gerstenmaier JF, Molteni F, Abdulla S, Montgomery SM, Kaneko A, Björkman A (2007). Impact of artemisinin-based combination therapy and insecticide-treated nets on malaria burden in Zanzibar. *PLoS Medicine* **4**(11):e309

Bockarie MJ, Davies JB, Thomson MC, Morgan HG (1990). The transmission of onchocerciasis at a forest village in Sierra Leone. I. *Simulium damnosum s.l.* biting densities and infection with *Onchocerca volvulus* at five representative sites. *Annals of Tropical Medicine and Parasitology* **84**(6): 587–97

Bogreau H, Renaud F, Bouchiba H, Durand P, Ass, S-B, Henry M-C, Garnotel E, Pradines B, Fusai T, Wade B, Adehossi E, Parola P, Kamil AM, Puijalon O, Rogier C (2006). Genetic diversity and structure of African *Plasmodium falciparum* populations in urban and rural areas. *The American Journal of Tropical Medicine and Hygiene* **74**(6): 953–9

Bousema T, Drakeley C, Gesase S, Hashim R, Magesa S, Mosha F, Otieno S, Carneiro I, Cox J, Msuya E, Kleinschmidt I, Maxwell C, Greenwood B, Riley E,

Sauerwein R, Chandramohan D, Gosling R (2010). Identification of hot spots of malaria transmission for targeted malaria control. *The Journal of Infectious Diseases* **201**(11): 1764–74

Bousema T, Griffin JT, Sauerwein RW, Smith DL, Churcher TS, Takken W, Ghani A, Drakeley C, Gosling R (2012). Hitting Hotspots: Spatial Targeting of Malaria for Control and Elimination. *PLoS Medicine* **9**(1):e1001165

Breman JG, de Quadros CA, Dowdle WR, Foege WH, Henderson DA, John TJ, Levine MM (2011). Laboratory investigations are indispensable to monitor the progress of measles elimination - results of the German Measles Sentinel 1999-2003. *PLoS Medicine* **8**(1): e1000405

Browning S (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* 178(4): 2123-32

Bruce-Chwatt LJ (1981). Alphonse Laveran's discovery 100 years ago and today's global fight against malaria. *Journal of the Royal Society of Medicine* **74**:531-36

Bruce MC, Galinski MR, Barnwell JW, Donnelly CA, Walmsley M, Alpers MP, Walliker D, Day KP (2000a). Genetic diversity and dynamics of *Plasmodium falciparum* and *P. vivax* populations in multiply infected children with asymptomatic malaria infections in Papua New Guinea. *Parasitology* **121** (3): 257–72

Bruce MC, Donnelly CA, Alpers MP, Galinski MR, Barnwell JW, Walliker D, Day, KP (2000b) Cross-species interactions between malaria parasites in humans. *Science* **287**: 845–848

Bruce MC, Macheso A, Galinski MR, Barnwell JW (2007). Characterization and application of multiple genetic markers for *Plasmodium malariae*. *Parasitology* **134**: 637–50

Campino S, Auburn S, Kivinen K, Zongo I, Ouedraogo, J.-B, Mangano V., Djimde A, Doumbo KO, Kiara MS, Nzila A, Borrman S, Marsh K, Michon P, Meller I, Siba P, Jiang H, Su XZ, Amaratunga C, Socheat D, Fairhurst MR, Imwong M, Anderson T, Nosten F, White JN, Gwilliam R, Deloukas P, MacInnis B, Newbold CI, Rockett, Clarke GT, Kwiatkowski DP (2011). Population Genetic Analysis of *Plasmodium falciparum* Parasites Using a Customized Illumina GoldenGate Genotyping Assay. *PloS One* **6**(6): e20251

Cappana E (2006). Grassi versus Ross: who solved the riddle of malaria? *International Microbiology* **9**(1):69-74

Carlton JM. Toward a malaria haplotype map (2007). *Nature and Genetics* **39**:5-6

Carter R, McGregor IA (1973). Enzyme variation in *Plasmodium falciparum* in the Gambia. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **67**(6): 830–837

Carter R, Voller A (1975). The distribution of enzyme variation in populations of Falciparum in Africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **69**(4): 371–376

Carter R, Graves PM (1988). Gametocytes. In Malaria Principles and Practice of Malariology. (ed . W. H. Wernsdorfer & I. McGregor), pp. 253-305. Edinburgh. Churchill Livingstone

Carter R, Mendis KN, Roberts D (2000). Spatial targeting of interventions against malaria. *Bulletin of the World Health Organization* **78**(12): 1401–11

Carter R (2001). Transmission blocking malaria vaccines. *Vaccine* **19**:2309-2314

Carter R & Mendis KN (2002). Evolutionary and Historical Aspects of the Burden of Malaria. *Clinical Microbiology Reviews* **15**(4):564-594

Castelli F, Odolini S, Autino B, Foca E, Russo R (2010). Malaria Prophylaxis: A Comprehensive Review. *Pharmaceuticals* **3**(10): 3212–3239

Ceesay SJ, Casals-Pascual C, Erskine J, Anya SE, Duah NO, Fulford AJC, Sesay SSS, Abubakar I, Dunyo S, Sey O, Palmer A, Fofana M, Corrah T, Bojang KA, Whittle HC, Greenwood BM, Conway DJ (2008). Changes in malaria indices between 1999 and 2007 in The Gambia: a retrospective analysis. *Lancet* **372**(9649): 1545–54

Centers for Disease Control and Prevention (CDC). Malaria: Larval Control and Other Vector Control Interventions:
http://www.cdc.gov/malaria/malaria_worldwide/reduction/vector_control.html

Chandiwana SK,Woolhouse ME (1991). Heterogeneities in water contact patterns and the epidemiology of *Schistosoma haematobium*. *Parasitology* **103**(3): 363–70

Charlwood JD, Alecrim WA (1989). Capture-recapture studies with the South American malaria vector Anopheles darlingi, Root. *Annals of Tropical Medicine and Parasitology* **83**(6): 569–76

Cheesman S, Creasey A, Degnan K, Kooij T (2007). Validation of PyrosequencingTM for accurate and high throughput estimation of allele frequencies in malaria parasites. *Molecular and Biochemical Parasitology* **152**:213-219

Cheesman S, Tanabe K, Sawai H, O'Mahony E, Carter R (2009). Strain-specific immunity may drive adaptive polymorphism in the merozoite surface protein 1 of the rodent malaria parasite *Plasmodium chabaudi*. *Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* **9**(2): 248–55

Cheesman S, O'Mahony E, Pattaradilokrat S, Degnan K, Knott S, Carter R (2010). A single parasite gene determines strain-specific protective immunity against malaria:

the role of the merozoite surface protein I. *International journal for parasitology* **40**(8): 951–61

Clarke SE, Bøgh C, Brown RC, Pinder M, Walraven GE, Lindsay SW (2000). Do untreated bed nets protect against malaria? *Transactions of the Royal Society of Tropical Medicine and Hygiene* **95**(5):457-62

Coatney GR, Collins WE, Warren McW, Contacos PG (1971). The Primate Malarias, 2nd Edn. U.S. Department of Health, Education and Welfare, NIH, Bethesda, Maryland, USA. Cited in Singh *et al*., 2004

Cole-Tobian JL, Biasor M, King CL (2005). High complexity of *Plasmodium vivax* infections in Papua New Guinean children. *The American Journal of Tropical Medicine and Hygiene* **73**(3): 626–33

Collins WJ, Greenhouse B, Rosenthal P, Dorsey G (2006). The use of genotyping in antimalarial clinical trials: a systematic review of published studies from 1995-2005. *Malaria journal* **5** (122) doi:10.1186/1475-2875-5-122

Contamin H, Fandeur T, Bonnefoy S, Skouri F, Ntoumi F, Mercereau-Puijalon O, (1995). PCR typing of field isolates of *Plasmodium falciparum. Journal of clinical microbiology* **33**(4): 944–51

Conway DJ, Greenwood BM, McBride JS (1991). The epidemiology of multiple-clone *Plasmodium falciparum* infections in Gambian patients. *Parasitology* **103**(1):1–6

Conway DJ, McBride JS (1991). Population genetics of *Plasmodium falciparum* within a malaria hyperendemic area. *Parasitology* **103**(1): 7–16

Conway DJ, Roper C, Oduola AM, Arnot DE, Kremsner PG, Grobusch MP, Curtis CF, Greenwood BM (1999). High recombination rate in natural populations of *Plasmodium falciparum. Proceedings of the National Academy of Sciences of the United States of America* **96**(8): 4506-11

Conway DJ, Machado RL, Singh B, Dessert P, Mikes ZS, Povoa MM, Oduola AM, Roper C (2001). Extreme geographical fixation of variation in the *Plasmodium falciparum* gamete surface protein gene Pfs48/45 compared with microsatellite loci. *Molecular and Biochemical Parasitology* **115**(2): 145–56

Cowan GJM, Creasey, AM, Dhanasarnsombut K, Thomas AW, Remarque EJ, Cavanagh DR, (2011). A malaria vaccine based on the polymorphic block 2 region of MSP-1 that elicits a broad serotype-spanning immune response. *PloS One* **6**(10): p.e26616

Cox-Singh J, Davis, TME, Lee K-S, Shamsul SSG, Matusop A, Ratnam S, Rahman HA, Conway DJ, Singh B (2008a). *Plasmodium knowlesi* malaria in humans is

widely distributed and potentially life threatening. *Clinical Infectious Diseases* **46**(2): 165-71

Cox-Singh J, Singh B (2008b). Knowlesi malaria: newly emergent and of public health importance? *Trends in Parasitology* **24**(9): 406–10

Craig MH, Snow RW, le Sueur D (1999). A climate-based distribution model of malaria transmission in sub-Saharan Africa. *Parasitology Today* **15**(3): 105–11

Creasey A, Fenton B, Walker A, Thaithong S, Oliveira S, Mutambu S, Walliker, D (1990). Genetic diversity of *Plasmodium falciparum* shows geographical variation. *The American Journal of Tropical Medicine and Hygiene* **42**(5): 403–13

Cregan B, Mudge J, Fickus EW, Marek LF, Danesh D, Denny R, Shoemaker RC, Matthews BF, Jarvik T, Young ND (1999) Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. *Theoretical and Applied Genetics* **98**: 919-928

Daubersies P, Sallenave-Sales S, Magne S, Trape JF, Contamin H, Fandeur T, Rogier C, Mercereau-Puijalon O, Druilhe P (1996). Rapid turnover of *Plasmodium falciparum* populations in asymptomatic individuals living in a high transmission area. The American *Journal of Tropical Medicine and Hygiene* **54**(1): 18–26

Day KP, Koella JC, Nee S, Gupta S, Read AF (1992). Population genetics and dynamics of *Plasmodium falciparum*: an ecological view. *Parasitology* **104**: Suppl(1987):S35–52

de Meeûs T, McCoy KD, Prugnolle F, Chevillon C, Durand P, Hurtrez-Boussès S, Renaud F (2007). Population genetics and molecular epidemiology or how to "*débusquer la bête*". *Journal of molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* **7**(2): 308–32

Dondorp MA, Nosten F, Yi P, Das D, Phyo AP, Tarning J, Lwin KM, Ariey F Hanpithakpong W, Lee SJ, Ringwald P, Silamut K, Imwong M, Chotivanich K, Lim P, Herdman T, An SS, Yeung S, Singhasivanon P, Day NPJ, Lindegardh N, Socheat D, White NJ (2009). Artemisinin Resistance in *Plasmodium falciparum* Malaria. *New England Journal of Medicine* **361**(5):455-467

Doolan DL, Dobaño C, Baird JK (2009). Acquired immunity to malaria. *Clinical Microbiology Reviews* **22**(1):13–36

Double MC, Peakall R, Beck NR and Cockburn A (2005). Dispersal, philopatry and infidelity: dissecting local genetic structure in superb fairy-wrens (*Malurus cyaneus*). *Evolution* 59: 625-635

Drakeley C, Schellenberg D, Kihonda J, Sousa CA, Arez AP, Lopes D, Lines J, Mshinda H, Lengeler C, Schellenberg JA, Tanner M, Alonso P (2003). An estimation of the entomological inoculation rate for Ifakara: a semi-urban area in a region of

intense malaria transmission in Tanzania. *Tropical Medicine and International Health* **8**(9):767–774

Druilhe P, Daubersies P, Patarapotikul J, Gentil C, Chene L, Chongsuphajaisiddhi T, Duffy PE (2007). Plasmodium in the placenta: parasites, parity, protection, prevention and possibly preeclampsia. *Parasitology* 134(13):1877–81

Dyer M, Day KP (2000). Commitment to gametocytogenesis in *Plasmodium falciparum*. *Parasitology today* **16**(3):102–7

Edvinsson B, Dardé M-L, Pelloux H, Evengård B (2007). Rapid genotyping of *Toxoplasma gondii* by pyrosequencing. *Clinical Microbiology and Infection* **13**(4):424-9

Eichner M, Diebner HH, Molineaux L, Collins WE, Jeffery GM, Dietz K (2001). Genesis, sequestration and survival of *Plasmodium falciparum* gametocytes: parameter estimates from fitting a model to malaria therapy data. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **95**(5):497–501

Ellegren H (2004). Microsatellites: simple sequences with complex evolution. *Nature reviews. Genetics* **5**(6): 435–45

Engelbrecht F, Togel E, Beck HP, Enwezor F, Oettli A, Felger I (2000). Analysis of *Plasmodim falciparum* infections in a village in Northern Nigeria: determination of msp2 genotypes and parasite-specific IgG responses. *Acta Tropica* **74**:63-71

England R, Pettersson M (2005). Pyro Q-CpGTM: Quantitative analysis of methylation in multiple CpG sites by Pyrosequencing®. *Nature methods*: 10.1038/NMETH800

Excoffier L, Smouse PE and Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitocondrial DNA restriction sites. *Genetics* **131**: 479-491

Färnert A, Snounou G, Rooth I, Bjorkman A (1997). Daily dynamics of *Plasmodium falciparum* subpopulations in a holoendemic area. *The American Journal of Tropical Medicine and Hygiene* **56**(5):538–547

Färnert A, Arez AP, Babiker HA, Beck HP, Benito A, Bjiirkman A, Rubio JM (2001). Genotyping of *Plasmodium falciparum* infections by PCR : a comparative multicentre study *Transactions of the Royal Society of Tropical Medicine and Hygiene* **95**: 225–232

Färnert A, Lebbad M, Faraja L, Rooth I (2008). Extensive dynamics of *Plasmodium falciparum* densities, stages and genotyping profiles. *Malaria journal* **7**:241 doi: 10.1186/1475-2875-7-241

Färnert A, Williams TN, Mwangi TW, Ehlin A, Fegan G, Macharia A, Lowe BS, Montgomery SM, Marsh K (2009). Transmission-dependent tolerance to multiclonal *Plasmodium falciparum* infection. *The Journal of Infectious Diseases* **200**(7): 1166–75.

Fegan GW, Noor AM, Akhwale WS, Cousens S, Snow RW (2007). Effect of expanded insecticide-treated bednet coverage on child survival in rural Kenya: a longitudinal study. *Lancet* **370**(9592):1035–9

Felger I, Smith T, Edoh D, Kitua A, Alonso P, Tanner M, Beck HP (1999). Multiple *Plasmodium falciparum* infections in Tanzanian infants. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **93** (Suppl 1): 29–34

Felger I, Tavul L, Kabintik S, Marshall V, Genton B, Alpers M, Beck HP (1994). *Plasmodium falciparum*: extensive polymorphism in merozoite surface antigen 2 alleles in an area with endemic malaria in Papua New Guinea. *Experimental Parasitology* **79**(2):106–16

Felsenstein J (2007). Theoretical evolutionary genetics. Department of Genome Sciences, University of Washington, Seattle Washington

Fillinger U, Sonye G, Killeen GF, Knols BGJ, Becker N (2004). The practical importance of permanent and semipermanent habitats for controlling aquatic stages of *Anopheles gambiae sensu lato* mosquitoes: operational observations from a rural town in western Kenya. *Tropical Medicine and International Health* **9**(12): 1274-89

França LTC, Carrilho E, Kist TBL (2002). A review of DNA sequencing techniques. *Quarterly reviews of biophysics* **35**(2): 169–200

Fraser DJ (2008). How well can captive breeding programs conserve biodiversity? A review of salmonids. *Evolutionary Applications*: doi:10.1111/j.1752-4571.2008.00036.x

Fried M, Duffy PE (1996). Adherence of *Plasmodium falciparum* to chondroitin sulphate A in the human placenta. *Science* **272**:1502-1504

Frischknecht F, Baldacci P, Martin B, Zimmer C, Thiberge S, Olivo-Marin J-C Shorte SL, Ménard R (2004). Imaging movement of malaria parasites during transmission by Anopheles mosquitoes. *Cellular Microbiology* **6**(7):687–94

Gadsby NJ, Carter R (2008). A Genetic Analysis of Two Strains of *Plasmodium chabaudi adami* that Differ in Growth and Pathogenicity. University of Edinburgh Research Archives: http://www.era.lib.ed.ac.uk/handle/1842/2760

Gamage-Mendis AC, Carter R, Mendis C, De Zoysa AP, Herath PR, Mendis KN (1991). Clustering of malaria infections within an endemic population: risk of malaria associated with the type of housing construction. *The American Journal of Tropical Medicine and Hygiene* **45**(1):77-85

Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, Paulsen IT, James K, Eisen JA., Rutherford K, Salzberg SL., Craig A, Kyes S, Chan M-S, Nene V, Shallom SJ, Suh B, Peterson J, Angiuoli S, Pertea M, Allen J, Selengut J, Haft D, Mather MW, Vaidya AB, Martin DMA, Fairlamb AH, Fraunholz MJ, Roos DS, Ralph SA, McFadden GI, Cummings LM, Subramanian G M, Mungall C, Venter JC, Carucci DJ, Hoffman SL, Newbold C, Davis RW, Fraser CM, Barrell B (2002). Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**:498-511

Ghebreyesus TA, Haile M, Witten KH, Getachew A, Yohannes AM, Yohannes M, Teklehaimanot HD (1999). Incidence of malaria among children living near dams in northern Ethiopia: community based incidence survey. *BMJ (Clinical research edition.)* **319**(7211): 663–6

Ghebreyesus TA, Haile M, Witten KH, Getachew A, Yohannes M, Lindsay SW, Byass P (2000). Household risk factors for malaria among children in the Ethiopian highlands. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **94**(1):17-21

Gilles HM (1993). Epidemiology of malaria. In: Bruce-Chwatt's Essential Malariology (H.M. Gilles and D.A. Warrell, eds), 3rd ed. pp. 124–163. London:

Goldstein DB, Clark AG (1995) Microsatellite variation in North American populations of *Drosophila melanogaster*. *Nucleic Acids Research* **23**:3882–3886

Goldstein DB, Schlötterer (eds) (1999) Microsatellites: evolution and applications. Oxford University Press, Oxford

Goodman SJ (1997). Rst Calc: a collection of computer programs for calculating estimates of genetic differentition from microsatellite data and a determining their significance. *Molecular Ecology* **6**: 881-885

Gorgas CW (1918). Sanitation in Panama. New York:D. Appleton and Company, 1918. 182-205

Gouagna LC, Bancone G, Yao F, Yameogo B, Dabiré KR, Costantini C, Simporé J, Ouedraogo JB, Modiano D (2010). Genetic variation in human HBB is associated with *Plasmodium falciparum* transmission. *Nature Genetics* 42(4): 328–31

Goudet J, Raymond M, Demeeus T, Rousset F (1996). Testing differentiation in diploid populations. *Genetics* **144**: 1933-1940

Goudet J (2001). FSTAT, a program to estimate and test gene diversities and fixation indices (version 2.9.3). Available from http://www.unil.ch/izea/softwares/fstat.html.

Government of Kenya (GoK): Poverty Reduction Strategy Paper for the period 2001-2004. Nairobi, Government of the Republic of Kenya, Ministry of Finance and Planning; 2001

Greenhouse B, Myrick A, Dokomajilar C, Woo JM, Carlson EJ, Rosenthal PJ, Dorsey G (2006). Validation of microsatellite markers for use in genotyping polyclonal *Plasmodium falciparum* infections. *The American Journal of Tropical Medicine and Hygiene* **75**(5): 836-42

Greenwood BM, Fidoc, DA, Kyle DE, Kappe SHI, Alonso PL, Collins FH, Duffy PE (2008). Malaria : progress, perils, and prospects for eradication. *The Journal of Clinical Investigation* **118**(4): 1266–1276

Groenen MAM, Megens HJ, Zare Y, Warren WC, Hillier LW, Crooijmans RPM, Vereijken A, Okimoto R, Muir WM, Cheng HH (2011). The development and characterization of a 60K SNP chip for chicken. *BMC Genomics* **12**(1): 274

Gunawardena DM, Wickremasinghe AR, Muthuwatta L, Weerasingha S, Rajakaruna J, Senanayaka T, Kotta PK, Attanayake N, Carter R, Mendis KN (1998). Malaria risk factors in an endemic region of Sri Lanka, and the impact and cost implications of risk factor-based interventions. *The American Journal of Tropical Medicine and Hygiene* **58**(5):533-542

Haasl RJ, Payseur BA (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity* **106**(1): 158–71

Hall N (2007). Advanced sequencing technologies and their wider impact in microbiology. *The Journal of experimental biology* **210**(9):1518–25
Hamad AA, El Hassan IM, El Khalifa AA, Ahmed GI, Abdelrahim SA, Theander TG, Arnot DE (2000). Chronic *Plasmodium falciparum* infections in an area of low intensity malaria transmission in the Sudan. *Parasitology* **120** (Pt 5):447–56

Hamilton BM (2009). DNA polymorphism (Chapter 8 – Molecular evolution). In Population Genetics (1st ed). Pp 235-282. Hoboken, New Jersey. Wiley-Blackwell

Haubold H, Hudson RR (2000). LIAN 3.0: detecting linkage disequilibrium in multilocus data. *Bioinformatics* **16**: 847-848

Hay SI, Omumbo JA, Craig MH, Snow RW (2000a). Earth observation, geographic information systems and *Plasmodium falciparum* malaria in sub-Saharan Africa. *Advances in Parasitology* **47**:173–215

Hay SI, Rogers DJ, Toomer JF, Snow RW (2000b). Annual *Plasmodium falciparum* entomological inoculation rates (EIR) across Africa: literature survey, internet access and review. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **94**: 113-127

Hay SI, Snow RW (2006). The malaria atlas project: Developing global maps of malaria risk. *PLoS Medicine* **3**(12):2204-2208

Hendrick PW (2005). A standardized genetic differentiation measure. *Evolution* **59** (8): 1633-38

Hill WG, Babiker HA (1995). Estimation of numbers of malaria clones in blood samples. *Proceedings of the Royal Society of Biological Sciences* **262**(1365): 249–57

Hindar K, Fleming I, Mcginnity P, Diserud O (2006). Genetic and ecological effects of salmon farming on wild salmon: modelling from experimental results. I*CES Journal of Marine Science* **63**(7): 1234–1247

Hughes AL (1992). Positive selection and interallelic recombination at the merozoite surface antigen-1 (MSA-1) locus of *Plasmodium falciparum*. *Molecular Biology and Evolution* **9**: 381–393

Hughes MK, Hughes AL (1995). Natural selection on Plasmodium surface proteins. *Molecular and Biochemical Parasitology* **71**: 99–113

Hughes AL, Vierra F (2001). Very large long-term effective population size in the virulent human malaria parasite *Plasmodium falciparum*. *Proceedings of the Royal Society of Biological science* **268**(1478):1855–60

Hurwitz, ES, Johnson D, Campbell, CC (1981). Resistance of *Plasmodium falciparum* malaria to sulfadoxine-pyrimethamine (Fansidar) in a refugee camp in Thailand. *Lancet* 1068-107

Iwagami M, Rivera PT, Villacorte EA, Escueta AD, Hatabu T, Kawazu S, Hayakawa T, Tanabe K, Kano S (2009). Genetic diversity and population structure of *Plasmodium falciparum* in the Philippines. *Malaria journal* **8**: 96. doi:10.1186/1475-2875-8-96

Jambou R, Martinelli A, Pinto J, Gribaldo S, Legrand E, Niang M, Kim N, Pharath L, Volnay B, Ekala MT, Bouchier T, Fandeur B, Berzosa F, Benito A, Ferreira DI, Vieira PP, Alecrim MG, Mercereau-Puijalon O, Cravo P (2010). Geographic structuring of the *Plasmodium falciparum* sarco(endo)plasmic reticulum Ca2+ ATPase (PfSERCA) gene diversity. *PloS One* 5(2): e9424. doi:10.1371/journal.pone.0009424

Jarne P, Lagoda PJ (1996). Microsatellites, from molecules to populations and back. *Trends in Ecology & Evolution* **11**(10): 424–9

Jensen JD, Kim Y, DuMont VB, Aquadro CF, Bustamante CD (2005). Distinguishing between selective sweeps and demography using DNA polymorphism data. *Genetics* **170**(3): 1401-10

Jongwutiwes S, Putaporntip C, Iwasaki T, Sata T, Kanbara H (2004). Naturally acquired *Plasmodium knowlesi* malaria in human, Thailand. *Emerging Infectious Diseases* **10**(12) 2211-3

Jiang H, Yi M, Mu J, Zhang L, Ivens A, Klimczak LJ, Huyen Y, Stephens RM, Su X (2008). Detection of genome-wide polymorphisms in the AT-rich *Plasmodium falciparum* genome using a high-density microarray. *BioMed Central Genomics* **9**:398-412

Jiang H, Li N, Gopalan V, Zilversmit MM, Varma S, Nagarajan V, Li J, Mu J, Hayton K, Henschen, B, Yi M, Stephens R, McVean G, Awadalla P, Wellems TE, Su X (2011). High recombination rates and hotspots in a *Plasmodium falciparum* genetic cross. *Genome Biology* **12** (4): R33

John TJ, Samuel R (2000). Herd immunity and herd effect: new insights and definitions. *European Journal of Epidemiology* **16**: 601-6

Jones MK, Good MF (2006). Malaria parasites up close. *Nature medicine* **12**(2):170–1

Joy DA, Feng X, Mu J, Furuya T, Chotivanich K, Krettli AU, Ho M, Wang A, White N, Suh E, Beerli P, Su X (2003). Early origin and recent expansion of *Plasmodium falciparum*. *Science* **300** (5617): 318-21

Juliano JJ, Porter K, Mwapasa V, Sem R, Roger, WO, Ariey F, Wongsrichanala C, Read A, Meshnick RS (2010a). Exposing malaria in-host diversity and estimating population diversity by capture-recapture using massively parallel Pyrosequencing. *PNAS* doi:10.1073/pnas

Juliano JJ, Gadalla N, Sutherland CJ, Meshnick SR (2010b). The perils of PCR: can we accurately "correct" antimalarial trials? *Trends in Parasitology* **26**(3): 119–24

Kaneko A (2010). A community-directed strategy for sustainable malaria elimination on islands: short-term MDA integrated with ITNs and robust surveillance. *Acta Tropica* **114**(3): 177–83

Karunaweera ND, Wijesekera SK, Wanasekera D, Mendis KN, Carter R (2003). The paroxysm of *Plasmodium vivax* malaria. *Trends in Parasitology* **19**(4):188-193

Kelly GC, Hii J, Batarii W, Donald W, Hale E, Nausien J, Pontifex S, Vallely A, Tanner M, Clements A (2010). Modern geographical reconnaissance of target populations in malaria elimination zones. *Malaria journal* **9**(1):289 doi:10.1186/1475-2875-9-289

Kimura M, Crow JF (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**: 725-38

Kimura M, Ohta T (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population. *Proceedings of the National Academy of Sciences USA* **75**: 2868-72

Kimura E, Mattei D, di Santi SM, Scherf A (1990). Genetic diversity in the major merozoite surface antigen of *Plasmodium falciparum*: high prevalence of a third polymorphic form detected in strains derived from malaria patients. *Gene* **91**(1): 57–62

Kleinschmidt I, Schwabe C, Benavente L, Torrez M, Ridl FC, Segura JL, Ehmer P, Nchama GN (2009). Marked increase in child survival after four years of intensive malaria control. *The American Journal of Tropical Medicine and Hygiene* **80**(6): 882–8

Knapik EW, Goodman A, Ekker M, Chevrette M, Delgado J, Neuhauss S, Shimoda N, Driever W, Fishman MC, Jacob HJ (1998) A microsatellite genetic linkage map for zebrafish (Danio rerio). *Nature Genetics* **18**: 338-343

Koepfli C, Schoepflin S, Bretscher M, Lin E, Kiniboro B, Zimmerman PA, Siba P, Smith TA, Mueller I, Felger I (2011). How much remains undetected? Probability of molecular detection of human Plasmodia in the field. *PloS One* **6**(4): e19010

Kolakovich KA, Ssengoba A, Wojcik K, Tsuboi T, al-Yaman F, Alpers M, Adams JH (1996). *Plasmodium vivax*: favored gene frequencies of the merozoite surface protein-1 and the multiplicity of infection in a malaria endemic region. *Experimental Parasitology* **83**: 11–18

Konate L, Zwetyengal J, Christopher R, Bischoff E, Fontenile D, Tall A, Spiegelz A,Kreuels B, Kobbe R, Adjei S, Kreuzberg C, von Reden C, Bäter K, Klug S, Busch W,Kun JFJ, Missinou MA, Lell B, Sovric M, Knoop H, Bojowald B, Dangelmaier O, Kremsner, Peter G (2002). New emerging *Plasmodium falciparum* genotypes in children during the transition phase from asymptomatic parasitemia to malaria. *The American Journal of Tropical Medicine and Hygiene* **66**(6): 653–8

Kozarewa I, Ning Z, Quail MA, Sanders MJ, Turner DJ (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. *Nature Methods* **6**(4): 291-295

Kumar S, Banks TW, Cloutier S (2012). SNP Discovery through Next-Generation Sequencing and Its Applications. *International Journal of Plant Genomics* 2012:1-15

Kyes S, Harding R, Black G, Craig A, Peshu N, Newbold C, Marsh K (1997). Limited spatial clustering of individual *Plasmodium falciparum* alleles in field isolates from coastal Kenya. *American Journal of Tropical Medicine and Hygiene* **57**:205-215

Lacroix R, Mukabana WR, Gouagna LC, Koella JC (2005). Malaria infection increases attractiveness of humans to mosquitoes. *PLoS Biology* **3**(9):.e298

Langhorne J, Ndungu FM, Sponaas AM, Marsh K (2008). Immunity to malaria: more questions than answers. *Nature Immunology* **9**(7): 725-32

Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, DeWinter AD, Berka J, Weiner M, Rothberg JM, Lohman KL (2003). A massively parallel PicoTiter Plate based platform for discrete picoliter-scale polymerase chain reactions. *Electrophoresis* **24**(21): 3769–77

Leberg PL (2002). Estimating allelic richness?: Effects of sample size. *Molecular Ecology* **11**: 2445–2449

Levinson G, Gutman GA (1987) High frequencies of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. *Nucleic Acids Research* **15**: 5323–5338

Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. (2002). Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular ecology* 11(12): 2453–65

Lindsay SW, Armstrong Schellenberg JRM, HA Zeiler, Daly RJ, Salum FM, Wilkins HA (1995). Exposure of Gambian children to *Anopheles gambiae* malaria vectors in an irrigated rice production area. *Medical and Veterinary Entomology* **9**:50-58

Liu S, Mu J, Jiang H, Su Xin-zhuan L (2008). Effects of Plasmodium falciparum mixed infections on in vitro antimalarial drug tests and genotyping. *The American journal of Tropical Medicine and Hygiene* **79**(2):178–84

Macdonald G (1957). The epidemiology and control of malaria. London, New York, Toronto, Oxford University Press`

Machado RLD, Povoa MM, Calvosa VSP, Ferreira MU, Rossit ARB, dos Santos EJ M, Conway DJ (2004). Genetic structure of *Plasmodium falciparum* populations in the Brazilian Amazon region. *The Journal of Infectious Diseases* **190**(9): 1547–55

Magesa SM, Mdira KY, Babiker HA, Alifrangis M, Färnert A, Simonsen PE, Bygbjerg IC, Walliker D, Jakobsen PH (2002). Diversity of *Plasmodium falciparum* clones infecting children living in a holoendemic area in north-eastern Tanzania. *Acta Tropica* **84**(2): 83–92

Malaria Vaccine Technology Roadmap (2004). (www.MalariaVaccineRoadmap.net)

Manske M, Miotto O, Campino S, Auburn S, Almagro-Garcia J, Maslen G, O'Brien J, Djimde A, Doumbo O, Zongo I, Ouedraogo JB, Michon P, Mueller I, Siba P,Nzila A, Borrmann S, Kiara SM, Marsh K, Jiang H, Su XZ, Amaratunga C, Fairhurst R, Socheat D, Nosten F, Imwong M, White NJ, Sanders M, Anastasi E, Alcock D,

Drury E, Oyola S, Quail MA, Turner DJ, Ruano-Rubio V, Jyothi D, Amenga-Etego L, Hubbart C, Jeffreys A, Rowlands K, Sutherland C, Roper C, Mangano V,Modiano D, Tan JC, Ferdig MT, Amambua-Ngwa A, Conway DJ, Takala-Harrison S, Plowe CV, Rayner JC, Rockett KA, Clark TG, Newbold CI, Berriman M,MacInnis B, Kwiatkowski DP. (2012). Analysis of *Plasmodium falciparum* diversity in natural infections by deep sequencing. *Nature*: **487**(7407):375-9

Mantel N (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**: 209-220

Mardis ER (2008). The impact of next-generation sequencing on genetics. *Trends in Genetics* **24**(3): 133-41

Mariette S, Chagné D, Lézier C, Pastuszka P, Raffin A, Plomion C, Kremer A (2001). Genetic diversity within and among *Pinus pinaster* populations: comparison between AFLP and microsatellite markers. *Heredity* **86**(Pt 4): 469–79

Mariette S, Le Corre V, Austerlitz, F, Kremer A (2002). Sampling within the genome for measuring within-population diversity: trade-offs between markers. *Molecular Ecology* **11**(7): 1145–56

Marshall VM, Anthony RL, Bangs MJ, Anders RF, Coppel RL (1994). Allelic variants of the *Plasmodium falciparum* merozoite surface antigen 2 (MSA-2) in a geographically restricted area of Irian Jaya. *Molecular and Biochemical Parasitology* **63**(1): 13–21

Matocq MD, Villablanca FX (2001). Low genetic diversity in an endangered species: recent or historic pattern? *Biological Conservation* **98**(1): 61–68

Mbogo CN, Snow RW, Khamala CP, Kabiru EW, Ouma JH, Githure JI, Marsh K, Beier JC (1995). Relationships between *Plasmodium falciparum* transmission by vector populations and the incidence of severe disease at nine sites on the Kenyan coast. *American Journal of Tropical Medicine and Hygiene* **52**:201-206

McBride JS, Walliker D, Morgan G (1982). Antigenic diversity in the human malaria parasite *Plasmodium falciparum*. *Science* **217**(4556): 254–7

Meirmans PG. (2006). Using the AMOVA framework to estimate a standardized genetic differentiation measure. *Evolution* **60**(11): 2399–2402

Mellouk S, Langsley G (1998). A primary malarial infection is composed of a very wide range of genetically diverse but related parasites. *The Journal of Clinical Investigation* **101**(9): 2008–16

Mendis KN, David PH, Carter R (1991). Antigenic polymorphism in malaria: is it an important mechanism for immune evasion? *Immunology Today* **12**(3): A34–7.

Metzker ML (2005). Emerging technologies in DNA sequencing. *Genome Research* **15**(12): 1767–76

Metzker ML (2010). Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**(1): 31–46

Michalakis Y, Veuille M (1996) Length variation of CAG/CAA trinucleotide repeats in natural populations of Drosophila melanogaster and its relation to the recombination rate. *Genetics* **143**: 1713–1725

Midega JT, Smith DL, Olotu A, Mwangangi JM, Nzovu JG, Wambua J, Nyangweso G, Mbogo M, Christophides GK, Marsh K, Bejon P, (2012). Wind direction and proximity to larval sites determines malaria risk in Kilifi District in Kenya. *Nature Communications* **3**: 674 doi:10.1038/ncomms1672

Miller LH, Good MF, Millon G (1994). Malaria pathogenesis. *Science* **264**:1878-1883

Minakawa N, Mutero CM, Githure JI, Beier JC, Yan G (1999). Spatial distribution and habitat characterization of anopheline mosquito larvae in Western Kenya. *The American Journal of Tropical Medicine and Hygiene* **61**(6):1010–6

Minakawa N, Sonye G, Dida GO, Futami K, Kaneko S (2008). Recent reduction in the water level of Lake Victoria has created more habitats for *Anopheles funestus*. *Malaria Journal* 7:119

Miura K, Keister DB, Muratova OV, Sattabongkot J, Long CA, Saul A (2007). Transmission-blocking activity induced by malaria vaccine candidates Pfs25/Pvs25 is a direct and predictable function of antibody titer. *Malaria Journal* **6**: 107

Mobegi VA, Loua KM, Ahouidi AD, Satoguina J, Nwakanma DC, Amambua-Ngwa, A, Conway DJ (2012). Population genetic structure of *Plasmodium falciparum* across a region of diverse endemicity in West Africa. *Malaria journal* **11**(1): 223. doi:10.1186/1475-2875-11-223

Mooney HA, Cleland EE (2001). The evolutionary impact of invasive species. *Proceedings of the National Academy of Sciences of the United States of America* **98**(10): 5446–51

Morin PA, Luikart G, Wayne RK (2004). SNPs in ecology, evolution and conservation. *Trends in Ecology & Evolution*. doi:10.1016/j.tree.2004.01.009

Mu J, Duan J, Joy DA, Su X (2002). Chromosome-wide SNPs reveal an ancient origin for *Plasmodium falciparum*. *Nature* **418**:323–326

Mu J, Awadalla P, Duan J, McGee KM, Joy DA, McVean GAT, Su X (2005). Recombination hotspots and population structure in *Plasmodium falciparum*. *PLoS Biology* **3**(10):p.e335

Mu J Awadalla, P Duan, J, McGee KM, Keebler J, Seydel K, McVean GAT, Su X (2007). Genome-wide variation and identification of vaccine targets in the *Plasmodium falciparum* genome. *Nature Genetics* **3**(1) 126–30

Mukabana WR, Takken W, Coe R, Knols BG (2002). Host-specific cues cause differential attractiveness of Kenyan men to the African malaria vector *Anopheles gambiae*. *Malaria Journal* **1**:17

Mulders MN, Truong AT, Muller CP (2001). Monitoring of measles elimination using molecular epidemiology. *Vaccine* **19** (17-19): 2245-9

Muller DA, Charlwood JD, Felger I, Ferreira C, Smith T (2001). Prospective risk of morbidity in relation to multiplicity of infection with *Plasmodium falciparum* in Sao Tome. *Acta Tropica* **78**:155-162

Mutuku FM, Alaii JA, Bayoh MN, Gimnig JE, Vulule JM, Walker ED, Kabiru E, Hawley WA (2006). Distribution, description, and local knowledge of larval habitats of *Anopheles gambiae* s.l. in a village in western Kenya. *The American Journal of Tropical Medicine and Hygiene* **74**(1): 44–53

Mwangi TW, Fegan G, Williams TN, Kinyanjui SM, Snow RW, Marsh K (2008). Evidence for over-dispersion in the distribution of clinical malaria episodes in children. *PloS One* 3(5): p.e2196

Myers SR, Griffiths RC (2003). Bounds on the minimum number of recombination events in a sample history. *Genetics* **163**(1): 375–94

Myrick A, Leemann E, Dokomajilar C, Hopkins H, Dorsey G, Kamya MR, Rosenthal PJ (2006). Short report: Dynamics of *Plasmodium falciparum* malaria after sub-optimal therapy in Uganda. *The American Journal of Tropical Medicine and Hygiene* **74**(5):758–61

National Vector Borne Disease Control Programme (NVBDCP). Guidelines on the use of larvivorous fish for vector control. Available at http://nvbdcp.gov.in/Doc/Guidelines-larvivorous-fish.pdf

Naylor R, Hindar K, Fleming Ian A, Goldburg R, Williams S, Volpe J, Whoriskey F, Eagle J, Kelso D, Mangel M (2005). Fugitive Salmon?: Assessing the Risks of Escaped Fish from Net-Pen Aquaculture. *Bioscience* **55**(5): 427-37

Nei M (1973). Analysis of genetic diversity in subdivided populations. *Proceedings of the National Academy of Sciences USA* **70**: 3321-3

Newbold C, Warn P, Black G, Berendt A, Craig A, Snow B, Msobo M, Peshu N, Marsh K (1997). Receptor-specific adhesion and clinical disease in *Plasmodium falciparum*. *American Journal of Tropical Medicine* **57**:389-398

Nkhoma SC, Nair S, Cheeseman IH, Rohr-Allegrini C, Singlam S, Nosten F, Anderson TJC (2012). Close kinship within multiple-genotype malaria parasite infections. *Proceedings of the Royal Society of Biological sciences* doi:10.1098/rspb.2012.0113

Noedl H (2008). Evidence of Artemisinin-Resistant Malaria in Western Cambodia. *New England Journal of Medicine* **359**(24):2619-2620

Nsobya SL, Kiggundu M, Joloba M, Dorsey G, Rosenthal PJ(2008). Complexity of Plasmodium falciparum clinical samples from Uganda during short-term culture. *The Journal of Infectious Diseases* **198**(10):1554–7

Ntoumi F, Contamin H, Rogier C, Bonnefoy S, Trape JF, Mercereau-Puijalon O (1995). Age-dependent carriage of multiple *Plasmodium falciparum* merozoite surface antigen-2 alleles in asymptomatic malaria infections. *American Journal of Tropical Medicine and Hygiene* **52**:81-88

Ntoumi F, Mercereau-Puijalon O, Ossari S, Luty A, Reltien J, Georges A, Millet P (1997). Plasmodium falciparum: sickle-cell trait is associated with higher prevalence of multiple infections in Gabonese children with asymptomatic infections. *Experimental Parasitology* **87**(1):39–46

O' Donell TW, Warren ST (2002). A decade of molecular studies of fragile X syndrome. *Annual Review of Neuroscience* **25**: 315-38

Olotu, A, Moris P, Mwacharo J, Vekemans J, Kimani D, Janssens M, Kai O, Jongert E, Lievens M, Leach A, Villafana T, Savarese B, Marsh K, Cohen J, Bejon P (2011). Circumsporozoite-specific T cell responses in children vaccinated with RTS,S/AS01E and protection against *P. falciparum* clinical malaria. *PloS One*: **6**(10): e25786

O'Meara WP, Bejon P, Mwangi TW, Okiro EA, Peshu N, Snow RW, Newton CRJC, Marsh K (2008). Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya. *Lancet* **372**(9649): 1555–62

Otten M, Aregawi M, Were W, Karema C, Medin A, Bekele W, Jima D, Gausi K, Komatsu R, Korenromp E , Low-Beer D, Grabowsky M (2009). Initial evidence of reduction of malaria cases and deaths in Rwanda and Ethiopia due to rapid scale-up of malaria prevention and treatment. *Malaria Journal* **8**: 14

Owusu-Agyei- S, Smith T, Beck H-P, Amenga-Etego L, Felger I (2002). Molecular Epidemiology of *Plasmodium falciparum* infections among asymptomatic inhabitants of a holoendemic malarious area in northern Ghana. *Tropical Medicine and International Health* **7**(5):421-428

Pasternak ND, Dzikowski, R (2009). PfEMP1: an antigen that plays a key role in the pathogenicity and immune evasion of the malaria parasite *Plasmodium falciparum*. *The international Journal of Biochemistry & Cell Biology* **41**(7): 1463-6

Paul RE, Packer MJ, Walmsley M, Lagog M, Ranford-Cartwright LC, Paru R, Day KP (1995). Mating patterns in malaria parasite populations of Papua New Guinea. *Science* **269**(5231): 709–11

Paulitschke M & Nash GB (1993). Membrane rigidity of red blood cells parasitized by different strains of *Plasmodium falciparum*. *The Journal of Laboratory and Clinical Medicine* **122**(5)581-589

Peakall R, Ruibal M and Lindenmayer DB (2003) Spatial autocorrelation analysis offers new insights into gene flow in the Australian bush rat, *Rattus fuscipes*. *Evolution* **57**: 1182-1195

Peakall R, Smouse PE (2006). Genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* **6**(1): 288–295

Peakall R, Smouse PE (2012) GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research - an update. *Bioinformatics*: doi:10.1093

Petit E, Balloux F, Goudet J (2001). Sex biased dispersal in a migratory bat : a characterization using sex-specific demographic parameters. *Evolution* **55**: 635-640

Pluess B, Tanser FC, Lengeler C, Sharp BL (2010). Indoor residual spraying for preventing malaria. *Cochrane Database of Systematic Reviews* **4**:.CD006657

Pumpaibool T, Arnathau C, Durand P, Kanchanakhan N, Siripoon N, Suegorn A, Sitthi-Amorn C, Renaud F, Harnyuttanakorn P (2009). Genetic diversity and population structure of *Plasmodium falciparum* in Thailand, a low transmission country. *Malaria Journal* **8**: 155. doi:10.1186/1475-2875-8-155

Ramírez-Soriano A, Ramos-O SE, Rozas J, Calafell F, Navarro A (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* **179**(1): 555-67

Reid H, Vallely A, Taleo G, Tatem AJ, Kelly G, Riley I, Harris I, Henri I, Iamaher S, Clements AC. (2010). Baseline spatial distribution of malaria prior to an elimination programme in Vanuatu. *Malaria Journal* **9**: 150 doi:10.1186/1475-2875-9-150

Ridley M (2004). Random events in population genetics. In Evolution, 3rd Ed., pp. 138-140. Blackwell Publishing company, Oxford, UK

Ronaghi M, Karamohamed S, Petterson B, Uhlen M, Nyren P (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* **242**:84-89

Ross R (1911). The prevention of Malaria. John Murray, London

Rousset F (1996). Equilibrium Values of Measures of Population Subdivision for Stepwise Mutation Processes. *Genetics* **142**: 1357-1362

Rozas J (2009). DNA Sequence Polymorphism Analysis using DnaSP. Pp. 337-350. In Posada, D. (ed.) Bioinformatics for DNA Sequence Analysis; Methods. In Molecular Biology Series Vol. 537. Humana Press, NJ, USA

Sanderson A, Walliker D, Molez JF (1981). Enzyme typing of *Plasmodium falciparum* from African and some other Old World countries. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **75**(2):263–7

Sanger F, Nicklen S, Coulson AR (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, USA* **74**: 5463-5467

Schellenberg JR, Abdulla S, Nathan R, Mukasa O, Marchant TJ, Kikumbih N, Mushi AK, Mponda H, Minja H, Mshinda H, Tanner M, Lengeler C (2001). Effect of large-scale social marketing of insecticide-treated nets on child survival in rural Tanzania. *Lancet* **357**(9264): 1241–7

Schlötterer C. (2000). Evolutionary dynamics of microsatellite DNA. *Chromosoma* **109**(6): 365–371

Schultz L, Wapling J, Mueller I, Ntsuke PO, Senn N, Nale J, Kiniboro B, Buckee CO, Tavul L, Siba MP, Reeder CJ, Barry AE (2010). Multilocus haplotypes reveal variable levels of diversity and population structure of *Plasmodium falciparum* in Papua New Guinea, a region of intense perennial transmission. *Malaria Journal* **9**(1): 336. doi:10.1186/1475-2875-9-336

Selkoe KA, Toonen RJ (2006). Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* **9**(5):615–29

Shauka AM, Breman JG, McKenzie FE (2010). Using the entomological inoculation rate to assess the impact of vector control on malaria parasite transmission and elimination. *Malaria Journal* **9**:122

Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler AD, Gibbs AR, Yu F (2010). A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Research* **20**(2): 273–80

Sievers AC, Lewey J, Musafiri P, Franke MF, Bucyibaruta BJ, Stulac SN, Rich ML,Karema C, Daily JP (2008). Reduced paediatric hospitalizations for malaria and febrile illness patterns following implementation of community-based malaria control programme in rural Rwanda. *Malaria Journal* **7**:167

Singh B, Kim SL, Matusop A, Radhakrishnan A, Shamsul SSG, Cox-Singh J, Thomas A, Conway DJ (2004). A large focus of naturally acquired *Plasmodium knowlesi* infections in human beings. *Lancet* **363**(9414):1017–24

Slatkin M (1995). A measure of population subdivision based on microsatellite allele frequency. *Genetics* **139**: 457-462

Smith DL, Dushoff J, Snow RW, Hay SI (2005). The entomological inoculation rate and *Plasmodium falciparum* infection in African children. *Nature* **438**(7067):492–5

Smith DL, McKenzie FE, Snow RW, Hay SI (2007). Revisiting the basic reproductive number for malaria and its implications for malaria control. *PLoS Biology*: **5**(3):.e42

Smouse PE, Long JC and Sokal RR (1986). Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology* **35**: 627-632

Smouse PE and Long JC (1992). Matrix correlation analysis in anthropology and genetics. *American Journal of Physical Anthropology* **35**: 187-213

Smouse PE and Peakall R (1999). Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity* **82**: 561-573

Smouse PE, Peakall R and Gonzales E (2008). A heterogeneity test for fine-scale genetic structure. *Molecular Ecology* **17**: 3389-3400

Sreekumar C, Hill DE, Miska KB, Vianna MCB, Yan L, Myers RL, Dubey JP (2005). Genotyping and detection of multiple infections of *Toxoplasma gondii* using Pyrosequencing. *International Journal for Parasitology* **35**(9): 991-9

Srinivas G, Edwin AR, Dhanraj B (2005). The use of personal protection measures against malaria in an urban population. *Public Health* **119**(5): 415–7

Staedke SG, Nottingham EW, Cox J, Kamya, MR, Rosenthal PJ, Dorsey G (2003). Short report: proximity to mosquito breeding sites as a risk factor for clinical malaria episodes in an urban cohort of Ugandan children. *The American Journal of Tropical Medicine and Hygiene* **69**(3): 244–6

Steketee RW (2009). Good news in malaria control... now what? *The American Journal of Tropical Medicine and Hygiene* **80**(6): 879–80

Stewart CN, Halfhill MD, & Warwick SI (2003). Transgene introgression from genetically modified crops to their wild relatives. *Nature reviews Genetics* **4**(10): 806–17

Su X-Z, Wellems TE, (1996). Towards a high resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats. *Genomics* **33**: 430-44

Su, X, Ferdig MT, Huang Y, Huynh CQ, Liu A, You J, Wootton, JC, Wellems TES (1999) A genetic map and recombination parameters of the human malaria parasite *P. falciparum*. *Science* **286**: 1351–1353

Su, X, Jiang H, Yi M, Mu J, Stephens RM (2009). Large-scale genotyping and genetic mapping in *Plasmodium* parasites. The Korean journal of parasitology 47(2):83-91

Suwanarusk R, Cooke BM, Dondorp AM, Silamut K, Sattabongkot J, White NJ, Udomsangpetch R (2011). *Blood* **117**(2): 381-92

Takala SL, Smith DL, Stine OC, Coulibaly D, Thera MA, Doumbo OK, Plowe CV (2006). A high-throughput method for quantifying alleles and haplotypes of the malaria vaccine candidate *Plasmodium falciparum* merozoite surface protein-1 19 kDa. *Malaria Journal* **5**: 31

Takken W, Charlwood JD, Billingsley PF GG (1998). Dispersal and survival of *Anopheles funestus* and *A . gambiae s . l* (Diptera?: Culicidae) during the rainy season in southeast Tanzania. *Bulletin of Entomological Research* 561–566.

Takken W, Knols BG (1999). Odor-mediated behavior of Afrotropical malaria mosquitoes. *Annual Review of Entomology* **44**: 131–57

Tanabe K, Sakihama N, Färnert A, Rooth I, Björkman A, Walliker D, Ranford-Cartwright L (2002). In vitro recombination during PCR of *Plasmodium falciparum* DNA: a potential pitfall in molecular population genetic analysis. *Molecular and Biochemical Parasitology* **122**(2): 211–6

Thaithong S, Sueblinwong T, Beale GH (1981). Enzyme typing of some isolates of Plasmodium falciparum from Thailand. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **75**(2): 268–70

The Bovine HapMap Consortium (2009). Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. *Science* **324**: 528-532

The malERA Consultative Group on Vaccines (2011) A Research Agenda for Malaria Eradication: Vaccines. *PLoS Medicine* **8**(1):S1-S10

Thomas CJ, Lindsay SW (2000). Local-scale variation in malaria infection amongst rural Gambian children estimated by satellite remote sensing. *Transactions of the Royal Society of Tropical Medicine* **94**: 159– 163

Tin F, Haling N, Lasserre, R (1982). Single-dose treatment of falciparum malaria with mefloquine: field studies with different doses in semi-immune adults and children in Burma. *Bulletin of the World Health Organization* **60**: 913-917

Tischer A, Santibanez S, Siedler A, Heider A, Hengel H (2004). Laboratory investigations are indispensable to monitor the progress of measles elimination--

results of the German Measles Sentinel 1999-2003. *Journal of Clinical Virology* **31**(3): 165-78

Trape JF, Lefebvre-Zante E, Legros F, Ndiaye G, Bouganali H, Druilhe P, Salem G (1992). Vector density gradients and the epidemiology of urban malaria in Dakar, Senegal. *The American Journal of Tropical Medicine and Hygiene* **47**(2):181–9

Trape JF, Rogier C, Konate L et al. (1994). The Dielmo project: al longitudinal study of natural malaria infection and the mechanisms of protective immunity in a community living in a holoendemic area of Senegal. *The American Journal of Tropical Medicine and Hygiene* **51**:123-137.

Trape J-F, Mercereau-Puijalon O (1999). Variation of *Plasmodium falciparum* msp1 block 2 and msp2 allele prevalence and of infection complexity in two neighbouring Senegalese villages with different transmission conditions. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **93**(1): 21–28

Trape, JF, Rogier C (2002). Combating malaria in Africa. *Trends in Parasitology* **18**(5): 224–30

Tsumori Y, Ndounga M, Sunahara T, Hayashida N, Inoue M, Nakazawa S, Casimiro, P, Rie I, Haruki U, Kazuyuki T, Osamu K, Culleton R (2011). *Plasmodium falciparum*: differential selection of drug resistance alleles in contiguous urban and peri-urban areas of Brazzaville, Republic of Congo. *PloS One* 6(8): e23430. doi:10.1371/journal.pone.0023430

Vafa M, Troye-Blomberg M, Anchang J, Garcia A, Migot-Nabias F (2008). Multiplicity of *Plasmodium falciparum* infection in asymptomatic children in Senegal: relation to transmission, age and erythrocyte variants. *Malaria Journal* **7**:17

Väli U, Einarsson A, Waits L, Ellegren H (2008). To what extent do microsatellite markers reflect genome-wide genetic diversity in natural populations? *Molecular Ecology* **17**(17): 3808–17

van der Hoek W, Konradsen F, Amerasinghe PH, Perera D, Piyaratne M, Amerasinghe FP (2003). Towards a risk map of malaria for Sri Lanka: the importance of house location relative to vector breeding sites. *International Journal of Epidemiology* **32**(2): 280–285

Vekemans X, Hardy OJ (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* **13**(4): 921–935

Vignal A, Milan D, San Cristobal MEA (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Review Literature And Arts Of The Americas* **34**: 275–305

Wacher TJ, Milligan PJ, Rawlings P, Snow WF (1994). Tsetse-trypanosomiasis challenge to village N'Dama cattle in The Gambia: field assessments of spatial and

temporal patterns of tsetse-cattle contact and the risk of trypanosomiasis infection. *Parasitology* **109** (Pt 2):149–62

Wahlgren M, Barragan A, Chen Q, Fernandez V, Hagblom P, Heddini A, Schlichtherle M, Scholander C, Sundström A, Treutiger CJ, Euler A, Carlson J (1998). Rosetting of malaria-infected erythrocytes: Ligands, host receptors and the involvement of serum proteins. In, Malaria: Parasite Biology, pathogenesis and protection pp. 387-398. *ASM Press, Washington DC*

Walliker D, Quakyi IA, Wellems TE, Mccutchan TF, Szarfman A, London WT, Corcoran LM, Burkot TR, Carter R (1987). Genetic analysis of the human malaria parasite *Plasmodium falciparum*. *Science* **236**:1661-1666

Walliker D (1994). The role of molecular genetics in field studies on malaria parasites. *International Journal for Parasitology* **24**(6):799–808

Walliker D, Babiker H, Ranford-Cartwright L (1998). The genetic structure of malaria parasite populations. In: Malaria: parasite biology, pathogenesis and protection, I. W Sherman ed. American Society for Microbiology. Washington D.C., 235-251

Wanji S, Mafo FF, Tendongfor N, Tanga MC, Tchuente E, Bilong Bilong CE, Njine T (2009). Spatial distribution, environmental and physicochemical characterization of *Anopheles* breeding sites in the Mount Cameroon region. *Journal of Vector Borne Diseases* **46**(1): 75-80

Watterson GA (1975). On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology* **7**:256-76

Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure. *Evolution* **38**: 1358-70

Wells TNC, Burrows JN, Baird JK (2010). Targeting the hypnozoite reservoir of *Plasmodium vivax*: the hidden obstacle to malaria elimination. *Trends in Parasitology* **26**(3):145–51

White NJ (2008). The role of anti-malarial drugs in eliminating malaria. *Malaria journal* **7** (Suppl 1): S8

White NJ (2011). A Vaccine for Malaria. *The New England Journal of Medicine* 10.1056/nejme1111777

WHO (2006). Guidelines for the Treatment of Malaria. Geneva: World Health Organization

WHO (2007). WHO Global Malaria Programme: Position Statement on ITNs. Available from: http://www.who.int/malaria/publications/atoz/itnspospaperfinal.pdf

WHO (2010). Guidelines for the treatment of malaria. Geneva: WHO. 1-210

WHO (2011): Malaria Vaccine Rainbow Tables.
[http://www.who.int/ vaccine_research/links/Rainbow/en/index.html]

WHO recommended insecticides for indoor residual spraying against malaria vectors. Available from: http://www.who.int/whopes/en/

Wilder JA, Hollocher H (2001). Mobile elements and the genesis of microsatellites in Dipterans. *Molecular Biology and Evolution* **18**:384-392

Wilkins HA, Blumenthal UJ, Hagan P, Hayes RJ, Tulloch S (1987). Resistance to reinfection after treatment of urinary schistosomiasis. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **81**(1):29–35

Williams TN (2006). Human red blood cell polymorphisms and malaria. *Current Opinion in Microbiology* **9**(4): 388–94

Williams SE, & Hoffman EA (2009). Minimizing genetic adaptation in captive breeding programs: A review. *Biological Conservation* **142**(11): 2388–2400

Wirth, DF (2002). Biological revelations. *Nature* **419**(6906): 495-496

Wongsrichanalai C, Pickard AL, Wernsdorfer WH, & Meshnick SR (2002). *Reviews Epidemiology of Drug-resistant Malaria* **2**: 209–218

Woolhouse MEJ, Dye C, Etard JF, Smith T, Charlwood JC, Garnett GP, Hagan P, Hii JLK, Ndhlovu PD, Quinnel RJ, Watts CH, Chandiwana SK, Anderson RM (1997). Heterogeneities in the transmission of infectious agents: Implications for the design of control programs. *Proceedings of the National Academy of Sciences* **94**(1):338-342

Wright S (1931). Evolution in Mendelian populations. Cited in Hamilton BM (2009). Population structure and gene flow (Chapter 4). In Population Genetics (1st ed). Pp 135-141. Hoboken, New Jersey. Wiley-Blackwell

Wright S (1943). Isolation by distance. *Genetics* **28**: 114-38

Wright S (1951). The genetic structure of populations. Cited in Hamilton BM (2009). Population structure and gene flow (Chapter 4). In Population Genetics (1st ed). Pp 135-141. Hoboken, New Jersey. Wiley-Blackwell

Wu Y, Ellis RD, Shaffer D, Fontes E, Malkin EM, Mahanty S, Fay MP, Narum D, Rausch K, Miles AP, Aebig J, Orcutt A, Muratova O, Song G, Lambert L, Zhu D, Miura K, Long C, Saul A, Miller LH, Durbin AP (2008). Phase 1 trial of malaria transmission blocking vaccine candidates Pfs25 and Pvs25 formulated with montanide ISA 51. *PloS One* **3**(7):.e2636

Yang Z, Miao J, Huang Y, Li X, Putaporntip C (2006). Genetic structures of geographically distinct *Plasmodium vivax* populations assessed by PCR/RFLP analysis of the merozoite surface protein 3-beta gene. *Acta Tropica* **100**: 205–12

Zhong D, Afrane Y, Githeko A, Yang Z, Cui L, Menge DM, Temu E, Yan G (2007). *Plasmodium falciparum* genetic diversity in western Kenya highlands. *The American Journal of Tropical Medicine and Hygiene* **77**(6): 1043–50

Zhou X, Ren L, Meng Q, Li Y, Yu Y, Yu J (2010). The next-generation sequencing technology and application. *Protein & Cell* **1**(6): 520–36

Zhou G, Githeko AK, Minakawa N, Yan G (2010). Community-wide benefits of targeted indoor residual spray for malaria control in the western Kenya highland. *Malaria Journal* **9**:67

APPENDIX A: STANDARD PCR PROTOCOL

Date: _____

REAGENT	WORKING CONC	WORKING VOLUME (50 µl)	X_____
10 µM Forward Primer	0.4 µM	2.0 µl	
10 µM Reverse Primer	0.4 µM	2.0 µl	
5 U/µl Immolase	1 U	0.2 µl	
25 mM dNTP mix	150 µM	0.3 µl	
50 mM $MgCl_2$	1.5 mM	1.5 µl	
10X PCR buffer	1X	5.0 µl	
10 µg/µl DNA		1.0 µl	
PCR H_20		38.0 µl	
TOTAL		**50.0** µl	

Conditions:

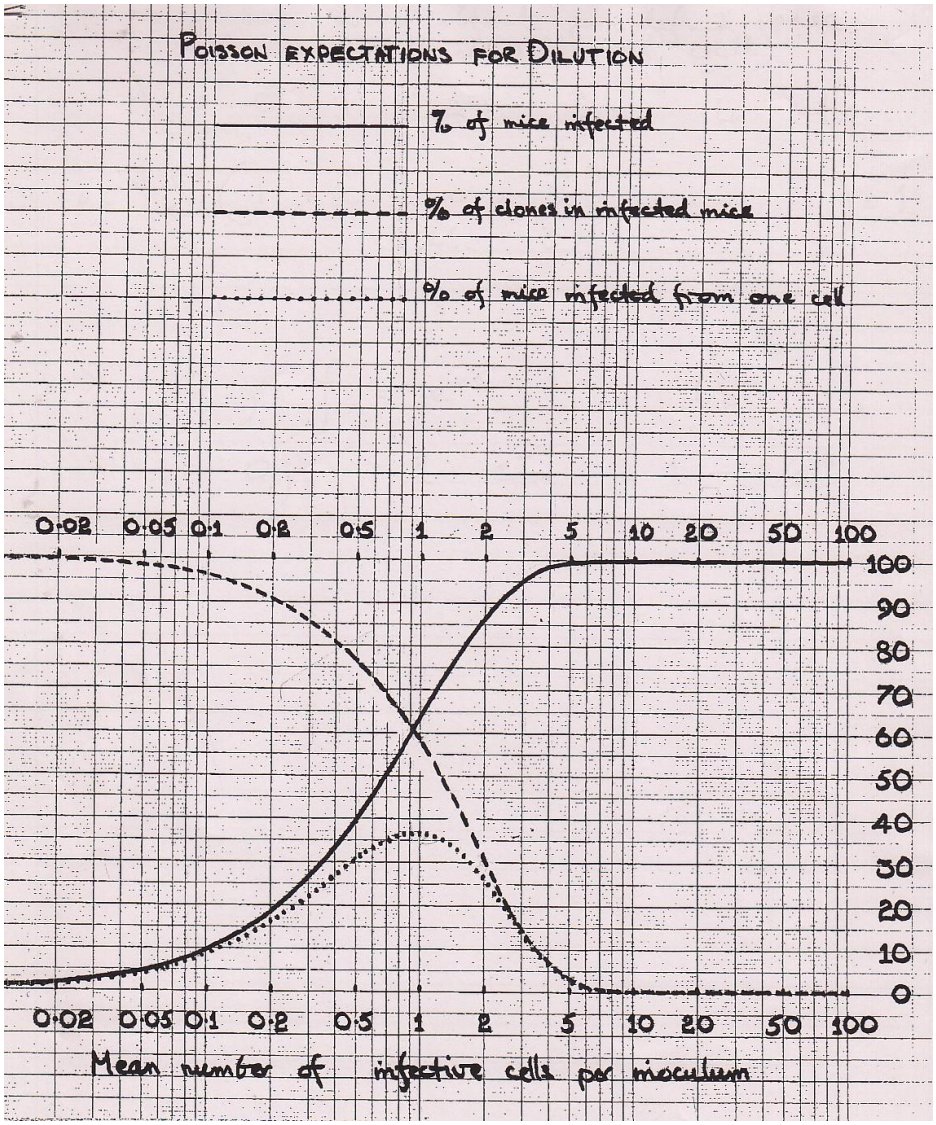
1st denaturation:	95 °c	7 min	
1st annealing	42 °c	1 min	
1st extension	72 °c	2 min	
2nd denaturation	95 °c	3 min	
2nd annealing	44 °c	1 min	
2nd extension	72 °c	2 min	
3rd denaturation	95 °c	1 min	X 38
3rd annealing	49 °c	1 min	
3rd extension	72 °c	2 min	
Final extension	72 °c	5 min	
	10 °c	4 ever	

Samples:

1.	11.	21.	31.	41.
2.	12.	22.	32.	42.
3.	13.	23.	33.	43.
4.	14.	24.	34.	44.
5.	15.	25.	35.	45.
6.	16.	26.	36.	46.
7.	17.	27.	37.	47.
8.	18.	28.	38.	48.
9.	19.	29.	39.	49.
10.	20.	30.	40.	50.

APPENDIX B: POISSON EXPECTATIONS FOR PARASITE DILUTIONS

This illustration serves as a guideline when cloning out parasites either in mice or blood cultures. As already described in **Section 5.2.1**, according to the Poisson expectation shown below ~80% positive wells are expected to occur in wells estimated to contain ~1.6 pRBCs and 99-100% positive wells in those estimated at ~5 pRBCS



APPENDIX C: KENYA ETHICAL REVIEW APPROVAL LETTER

KENYATTA NATIONAL HOSPITAL
Hospital Rd. along, Ngong Rd.
P.O. Box 20723, Nairobi.
Tel: 726300-9
Fax: 725272
Telegrams: MEDSUP", Nairobi.
Email: KNHplan@Ken.Healthnet.org

Ref: KNH-ERC/ 01/ 94

4th February 2008

Dr. Wolfgang Richard Mukabana
School of Biological Sciences
University of Nairobi

Dear Dr. Mukabana

Re Annual renewal for study titled "Active Monitoring of malaria prevalence and incidence to assess the impact of mosquito control interventions for malaria control on Rusinga Island" (P156/9/2005)

This is to grant you annual renewal for research proposal Ref. No.P156/9/2005.

The renewal periods are 11th April 2007 – 10th April 2008.
Yours sincerely

PROF A N GUANTAI
SECRETARY, KNH-ERC

c.c. Prof. K.M. Bhatt, Chairperson, KNH-ERC
The Deputy Director CS, KNH

APPENDIX D: MALI ETHICAL REVIEW APPROVAL LETTER

MINISTERE DE L'ENSEIGNEMENT
SUPERIEUR ET DE LA RECHERCHE SCIENTIFIQUE
› › › UNIVERSITE DE BAMAKO › › ›

FACULTE DE MEDECINE DE PHARMACIE ET
D'ODONTO-STOMATOLOGIE / BP 1805

: (223) 20 22 52 77 — : (223) 20 22 96 58

BAMAKO – MALI

Le Président du Comité d'Ethique
Prof. Mamadou Marouf KEITA

N°09 40 /FMPOS

Bamako le 28 mai 2009

A Monsieur le Docteur Mahamadou Soumana SISSOKO
Chercheur au Département d'Epidémiologie
des Affections Parasitaires (DEAP) Bamako

Cher Monsieur,

Le comité d'éthique institutionnel de la FMPOS a examiné en sa session plénière du 16 mai 2009, **la version 0.1 du 11 mai 2009** du protocole de l'étude intitulée: **«Détection des zones à risque élevé de transmission du paludisme dans les villages de Sotuba et Kollé en 2009»**, ainsi que **la version 0.1** du formulaire de consentement.

Le comité d'éthique institutionnel de la FMPOS constate que vous avez pris en compte ses commentaires à savoir :

1. ajouter le résumé du protocole,
2. préciser la compensation,
3. indiquer les éléments du budget,
4. la restitution des résultats de l'étude à la population et au comité d'éthique,
5. fournir le schéma du piège lumineux,
6. la prévision d'une visite de terrain de deux membres du comité d'éthique,
7. ajouter le curriculum vitae (CV) des nouveaux investigateurs,
8. ajouter le pourcentage du temps des investigateurs.

En conséquence, le comité d'éthique a décidé d'approuver **la version 0.1 du 11 mai** de votre protocole **et la version 0.1** du formulaire de consentement.

Le comité d'éthique vous autorise à entreprendre l'étude et vous souhaite plein succès.

LE PRESIDENT
DU COMITE D'ETHIQUE

Faculté de Médecine de Pharmacie
Le Président
Comité d'Ethique

Prof. Mamadou Marouf KEITA

APPENDIX E: CAMEROON ETHICAL REVIEW APPROVAL LETTER

REPUBLIQUE DU CAMEROUN
Paix - Travail – Patrie

MINISTERE DE LA SANTE PUBLIQUE

DELEGATION REGIONALE
POUR LE SUD OUEST
Tel 3332 22 10
Tel/Fax 3332 22 62
Email: pdph_sw@yahoo.com

REPUBLIC OF CAMEROON
Peace - Work - Fatherland

MINISTRY OF PUBLIC HEALTH

REGIONAL DELEGATION
FOR THE SOUTH WEST

Date: 2 2 JUIL 2009

THE REGIONAL DELEGATE

Ref R11 /MPH/SWP/PDPH/ FP-Research/ 527784

TO:

Dr. Anong Damian Nota
Department of Biochemistry and
Microbiology, Faculty of Science,
University of Buea

SUBJECT: **Ethical clearance and authorisation**
To collect specimens for Research

After a careful review of the study proposal presented to us by this group of researchers, we are sure that medical ethics will be respected as the collections will be effected by trained medical staff following internationally agreed guidelines after the consent of the individual patients has been obtained.

With the clarity of their methodology and view the importance of this research in future medical care of patients, we have no objection to them carrying out this research on *"The Genetic Diversity of P. falciparum isolates from Cameroon in Relation to Malaria severity"*.

We are therefore, inviting the concern institutions and communities to give the researchers all the necessary assistance they will need to enable them carry out this study of scientific importance.

The Regional Delegate of Public Health

Dr Chuwanga John Ndengue

www.minsante.cm/www.minsante.gov.cm

APPENDIX F: SUMMARY OF THE *P. FALCIPARUM* SAMPLES ANALYSED FROM CAMEROON, MALI AND KENYA.

The sample ID is given as C = Cameroon, H = Mali and K = Kenya. The a and b for example C3a and C3b represent the major clone and minor clone respectively. The columns indicated as letters A - M represent the SNP IDs with the chromosome SNP position as in **Chapter 3**: Two SNP positions were eliminated as they had missing data. The letters represent the SNPs: **A** - Pf-01-101502, **B** - Pf-01-205304, **C** - Pf-01-393153, **D** - Pf-01-411796, **E** - Pf-03-138-7, **F** - Pf-03-350224, **G** - Pf-03-613146, **H** - Pf-03-842383, **I** - Pf-03-967838, **J** - Pf-10-658467, **K** - Pf-10-992433, **L** - Pf-14-279668 & **M** - Pf-14-1710196. The numbers in the aforementioned columns represent 1 = A, 2 = C, 3 = G and 4 = T. The columns indicated as Poly α, TA81, TA17, TA53, TA40, PFPK2, TA60 & TA43 represent the 8 analysed microsatellite loci and the numbers in the column represent the allelic polymorphisms observed in each locus. The bold alleles represent the peak alleles and are classified as major clones in this case.

No.	Sample	Pop	Poly α	TA81	TA17	TA53	TA40	PFPK2	TA60	TA43	A	B	C	D	E	F	G	H	I	J	K	L	M
1	C2	Cameroon	177	164	213	142	0	246	257	0	1	2	2	1	3	4	2	1	1	1	1	1	4
2	C3a	Cameroon	180	182	247	159	0	237	232	0	1	2	2	1	4	4	2	1	1	1	1	1	1
3	C3b	Cameroon	180	182	247	159	0	237	232	0	1	4	1	1	4	4	2	1	1	1	1	4	4
4	C4a	Cameroon	168	164	**188**	139	0	267	229	0	1	2	2	1	4	4	2	1	1	1	3	1	4
5	C4b	Cameroon	168	164	212	139	0	267	229	0	1	4	2	1	4	4	2	1	1	1	1	1	4
6	C5a	Cameroon	165	161	**278**	142	0	243	238	0	1	4	2	1	4	4	2	1	1	1	1	1	4
7	C5b	Cameroon	165	161	258	142	0	243	238	0	1	4	2	1	4	4	2	1	1	1	1	1	4
8	C6a	Cameroon	**133**	**170**	216	142	0	249	245	0	1	4	1	1	4	4	2	1	1	1	1	1	4
9	C6b	Cameroon	165	176	216	142	0	249	245	0	1	2	1	1	4	4	2	1	1	1	1	1	4
10	C11	Cameroon	159	173	240	142	0	261	241	0	1	2	2	1	4	4	2	1	1	1	3	1	1
11	C14a	Cameroon	180	170	243	142	0	**246**	232	0	1	2	2	1	4	4	2	1	1	1	1	1	4
12	C14b	Cameroon	180	170	243	142	0	237	232	0	4	4	2	1	3	2	2	1	1	1	3	1	1
13	H1a	Mali	**162**	0	**213**	**150**	**267**	240	**322**	243	1	2	2	1	4	4	2	1	1	1	1	1	4
14	H1b	Mali	176	0	228	139	278	240	328	243	1	2	2	1	4	4	2	1	1	1	1	1	4
15	H2a	Mali	156	**176**	213	139	**267**	0	**302**	213	1	4	2	1	4	4	2	1	1	1	1	1	4
16	H2b	Mali	156	**176**	213	139	278	0	309	213	1	2	2	1	3	4	2	1	1	1	1	1	4
17	H3a	Mali	171	161	228	142	269	246	322	**238**	1	2	2	1	4	4	2	1	1	1	3	1	4
18	H3b	Mali	171	161	228	142	269	246	322	222	1	2	2	1	4	4	2	1	1	1	3	1	4
19	H4	Mali	184	170	258	146	269	240	322	232	1	2	2	1	3	1	2	1	1	1	1	1	4
20	H5	Mali	162	170	243	136	261	237	0	244	1	2	2	1	3	4	2	1	1	1	1	1	4
21	H6	Mali	0	176	228	142	287	243	309	213	1	2	2	1	3	4	2	1	1	1	3	1	4
22	H8	Mali	156	173	213	142	276	261	312	226	1	2	2	1	4	4	2	1	1	1	1	1	4
23	H12	Mali	168	176	243	139	276	240	322	213	1	2	2	1	4	4	2	1	1	1	1	1	4
24	H13	Mali	180	164	0	142	278	237	309	244	1	2	2	1	4	4	2	1	1	1	1	1	4
25	H16a	Mali	**168**	167	0	142	269	240	312	254	1	2	2	1	4	4	2	1	1	1	1	1	4

No.	Sample	Pop	Poly α	TA81	TA17	TA53	TA40	PFPK2	TA60	TA43	A	B	C	D	E	F	G	H	I	J	K	L	M
26	H16b	Mali	165	167	0	142	269	240	312	254	1	4	2	1	4	4	2	1	1	1	1	4	4
27	H18	Mali	184	176	198	142	287	0	322	0	2	4	2	1	4	4	2	1	1	3	1	1	4
28	H21a	Mali	168	176	198	142	249	243	306	244	0	2	2	1	4	4	2	1	1	1	3	1	4
29	H21b	Mali	168	176	198	142	249	243	306	244	0	2	2	1	4	4	2	1	1	1	3	1	4
30	H23	Mali	190	170	233	142	269	240	325	254	1	2	2	2	3	2	2	1	1	1	3	1	4
31	H24a	Mali	165	167	243	139	230	0	318	**244**	1	4	2	1	4	4	2	1	1	1	1	4	4
32	H24b	Mali	165	167	243	139	230	0	318	263	1	4	2	1	4	4	2	1	1	1	1	4	4
33	H26	Mali	208	176	228	142	267	237	0	238	1	2	2	1	4	2	2	1	1	1	1	1	4
34	H27a	Mali	162	**167**	**204**	142	**252**	**237**	**315**	257	1	2	2	1	4	4	2	1	1	1	1	1	4
35	H27b	Mali	0	170	183	142	261	240	328	257	1	2	2	1	4	4	2	1	1	3	3	1	4
36	H28	Mali	174	179	236	142	269	243	318	234	1	2	2	1	4	4	2	1	1	1	3	1	4
37	H30a	Mali	**171**	0	**228**	142	263	0	318	0	1	2	2	1	4	4	2	1	1	1	3	4	4
38	H30b	Mali	165	0	243	142	263	0	322	0	1	2	2	1	4	4	2	1	1	1	3	4	4
39	H31	Mali	0	0	243	0	267	264	315	226	1	4	2	1	4	2	2	1	1	3	3	1	4
40	H34	Mali	171	167	233	139	273	0	322	234	1	2	2	1	3	2	2	1	1	1	1	1	4
41	H35	Mali	159	170	213	0	269	243	309	241	1	2	2	1	4	4	2	1	1	1	1	1	4
42	H37	Mali	190	164	243	142	267	243	315	250	1	2	2	1	4	4	2	1	1	1	1	1	4
43	H38	Mali	190	164	243	142	267	243	315	250	1	2	2	1	4	4	2	1	1	1	1	1	4
44	H40a	Mali	162	167	201	142	269	**309**	0	232	1	4	2	1	4	4	2	1	1	1	3	1	4
45	H40b	Mali	162	167	201	142	269	**309**	0	232	1	4	2	1	4	4	2	1	1	1	2	1	4
46	H41a	Mali	190	170	213	142	230	255	318	**254**	1	2	2	1	4	2	2	1	1	1	1	1	4
47	H41b	Mali	190	170	213	142	230	255	318	234	1	2	2	1	4	2	2	1	1	1	1	1	4
48	H42a	Mali	159	167	**213**	**142**	**246**	**243**	**312**	219	1	2	2	1	3	4	2	1	1	1	1	1	4
49	H42b	Mali	159	167	198	146	273	264	315	219	1	2	2	1	4	2	2	1	1	1	3	4	4
50	H43	Mali	165	173	228	0	234	237	322	238	1	2	2	1	4	4	2	1	1	1	1	1	4

No.	Sample	Pop	Poly α	TA81	TA17	TA53	TA40	PFPK2	TA60	TA43	A	B	C	D	E	F	G	H	I	J	K	L	M
51	H44	Mali	0	170	228	0	273	246	322	232	1	2	2	1	3	4	2	1	1	1	1	1	4
52	H45	Mali	0	170	228	0	269	246	309	247	2	2	2	1	4	4	2	1	1	3	3	1	4
53	H46	Mali	165	164	228	142	267	237	322	257	1	2	2	1	4	4	2	1	1	1	3	1	4
54	H47a	Mali	174	167	0	0	0	243	0	0	1	4	2	1	4	4	2	1	1	1	3	4	4
55	H47b	Mali	174	167	0	0	0	243	0	0	1	1	2	1	4	4	2	1	1	1	3	4	4
56	H48a	Mali	0	167	0	142	0	249	312	0	1	2	2	1	3	4	2	1	1	1	1	1	4
57	H48b	Mali	0	167	0	142	0	249	312	0	2	2	2	1	3	4	2	1	1	1	1	1	4
58	H50a	Mali	**171**	**179**	**233**	**142**	**267**	0	309	257	1	2	2	1	4	4	2	1	1	1	3	4	4
59	H50b	Mali	168	176	228	136	269	0	0	0	1	2	2	1	4	4	2	1	1	1	1	1	4
60	H51a	Mali	180	167	228	142	269	246	309	210	2	2	2	1	4	4	2	1	1	1	0	1	4
61	H51b	Mali	180	167	228	142	269	246	309	210	1	1	2	1	4	4	2	1	1	1	0	1	4
62	K64a	Kenya	174	167	198	142	0	255	0	**247**	1	2	0	1	4	4	2	1	0	1	1	4	4
63	K64b	Kenya	174	167	198	142	0	255	0	222	1	2	0	1	1	4	2	1	0	1	3	4	4
64	K157	Kenya	184	167	213	142	0	237	0	238	1	4	0	1	4	4	2	1	1	1	1	4	4
65	K173	Kenya	156	164	243	146	0	243	0	244	1	2	0	1	4	4	2	1	1	1	3	1	4
66	K193a	Kenya	168	164	201	142	0	243	0	238	1	2	0	1	3	4	2	1	1	1	1	4	4
67	K193b	Kenya	168	164	201	142	0	243	0	238	1	4	0	1	4	2	2	1	1	3	3	1	4
68	K194a	Kenya	165	164	222	159	0	243	0	**244**	1	2	0	1	3	4	2	1	1	1	3	4	4
69	K194b	Kenya	165	164	222	159	0	243	0	225	1	2	0	1	1	4	2	1	1	1	3	1	4
70	K195	Kenya	190	164	228	0	0	240	0	248	1	4	0	1	4	2	2	1	1	1	3	4	4
71	K201a	Kenya	145	179	198	142	0	249	0	**251**	1	4	0	1	4	4	2	1	1	1	3	1	4
72	K201b	Kenya	145	179	198	142	0	249	0	232	1	4	0	1	4	4	2	1	1	1	3	1	4
73	K205	Kenya	156	164	228	142	0	255	0	**222**	1	4	0	1	3	4	2	1	1	1	1	1	4
74	K207a	Kenya	165	173	228	146	0	**240**	0	**204**	1	2	0	1	4	4	2	1	1	1	1	1	4
75	K207b	Kenya	165	173	228	146	0	249	0	232	1	4	0	1	4	4	2	1	1	1	1	4	4

No.	Sample	Pop	Poly α	TA81	TA17	TA53	TA40	PFPK2	TA60	TA43	A	B	C	D	E	F	G	H	I	J	K	L	M
76	K225	Kenya	**168**	**164**	198	142	0	**240**	0	**244**	1	4	0	1	4	4	2	1	1	1	1	1	4
77	K226a	Kenya	196	170	**228**	146	0	243	0	225	1	4	0	1	4	4	4	1	1	1	3	1	4
78	K226b	Kenya	196	170	243	146	0	243	0	225	1	4	0	1	4	4	4	1	1	1	1	4	4
79	K228a	Kenya	**162**	173	**243**	142	0	240	0	214	1	2	0	1	4	4	4	1	1	1	1	1	4
80	K228b	Kenya	165	173	228	142	0	240	0	214	1	2	0	1	1	4	4	1	1	1	1	4	4
81	K230a	Kenya	162	176	243	146	0	**240**	0	235	1	2	0	1	4	4	4	1	1	1	3	1	4
82	K230b	Kenya	162	176	243	146	0	243	0	235	1	2	0	1	4	2	4	1	1	3	3	1	4
83	K241a	Kenya	168	164	228	142	0	249	0	254	1	4	0	1	4	4	4	1	1	3	1	4	4
84	K241b	Kenya	168	164	228	142	0	249	0	254	1	4	0	1	1	4	4	1	1	3	1	4	4

APPENDIX G: FORMULAE

1. FSTAT computation of expected heterozygosity (H_{sk} or Hexp**)** using microsatellites data:

$$H_{sk} = \frac{n_k}{n_k - 1}\left(1 - \sum_{i=1}^{n} P_i^2\right)$$

Where: **H_{sk} or Hexp** = unbiased estimator of gene diversity or heterozygosity (see Nei, 1987, eq 7.39 p 164) per sample k

n_k = size of sample k (sample k is the subpopulations i.e. Cameroon, Kenya & Mali)

P_i = frequency of the ith allele

2. FSTAT determination of Nei's G_{ST} with the microsatellites data

The value **H_s** or **Hexp (s)** is given by the formula:

$$H_s = \frac{\tilde{n}}{\tilde{n} - 1}\left[1 - \sum_{i=k} P - \frac{H_o}{2\tilde{n}}\right]$$

Where: H_o = Observed heterozygosity but since this is a haploid, diploid genotypes were obtained by coding the same allele twice, thus:

$H_0 = 1 - \sum_i = \sum_k P_{i=k}$ (P = frequency of the ith or kth allele)

The sum of $\sum_i$ = $\sum_k$

$\tilde{n}$ is the harmonic mean of the nk (size of sample k). Harmonic mean of the sample size as the sample size varies in each of the analysed regions

Note that these are not weighted by sample size.

The value of Ht is given by the equation:

$$H_t = 1 - \sum_i + H_s[(\tilde{n}\, np)]/- H_o\,(2\tilde{n}\ np)$$

Where: np = number of samples

To generate G_{ST}', the formula is:

$$\frac{H_t' - H_s}{H_t'}$$

Where: H_t' is the overall gene diversity independent of sample size and H_s is as shown above.

3. Analysis of molecular variance (AMOVA) by GenAlEx

The formulae for AMOVA are adapted from those of the analysis of variance (ANOVA) except the sums of squares (SS) and mean sums of squares (MS) are obtained from a genetic distance matrix.

$$SS_{TOT} = \frac{\sum d^2{}_{ij}}{2N}$$

Where: SS_{TOT} = total sums of squares

d^2_{ij} = the squared genetic distance between the *i*th and *j*th of different alleles
N = Total number of samples analysed

$$SS_{WP1} = \frac{\sum d^2{}_{ij}}{2n_1}$$

Where: SS_{WP1} = Sum of squares within population 1

d^2_{ij} = the squared genetic distance between the *i*th and *j*th alleles

n = number of samples within population 1

The total sums of squares within all populations (SS_{WP}) is given by $SS_{WP1} + SS_{WP2}$, SS_{WP3}……SS_{WPn}

The sums of squares among populations (SS_{AP}) is given by: $SS_{TOT} - SS_{WP}$

The mean sums of squares within populations (MS_{WP}) are computed as:

$$MS_{WP} = \frac{SS_{WP}}{df_{WP}}$$

Where: SS_{WP} = Sums of squares within populations

df_{WP} = Degrees of freedom within populations N - N_p; N is the number of populations NP is number of populations with kth populations explained later.

The mean sums of squares among populations (MS_{AP}) are computed as:

$$MS_{AP} = \frac{SS_{AP}}{df_{AP}}$$

Where: SS_{AP} = Sums of squares among populations

Df_{AP} = Degrees of freedom among populations = $N_p - 1$; NP is number of samples with kth populations explained later.

Variance estimates are then given by:

Variance within populations (V_{WP}) = MS_{WP}

Variance among populations (V_{AP}) is:

$$V_{AP} = \frac{MS_{AP} - MS_{WP}}{N_0}$$

Where: $N_0 = \frac{1}{(N_p - 1)} \times \left(\sum_{k=1}^{N_p} n_{p_k} - \left(\frac{\sum_{k=1}^{N_p} {n_p}^2_k}{\sum_{k=1}^{N_p} n_{p_k}} \right) \right)$

Where: N = Number of samples

N_p = Number of populations with kth populations

N_k = Number of samples in the kth populations

The estimation of the variance among populations (V_{AP}) relative to the total variance (ϕ_{PT}) is finally computed as:

$$\emptyset_{PT} = \frac{V_{AP}}{V_{AP} + V_{WP}}$$

4. Mantel test analysis by GenAlEx

GenAlEx performs a Mantel test using the formulae (Smouse *et al*. 1986; Smouse & Long 1992):

$$SS_x = \sum_{i \neq j}^{N} (x_{ij} - \bar{x})^2$$

Where: SS_x = sum of squares for the x matrix

N = total number of samples in the x matrix

i & j = alleles represented in the population

$\sum(x_{ij} - \bar{x})^2$ = The square of how much each value in the x matrix deviates from the mean

$$SS_y = \sum_{i \neq j}^{N} (y_{ij} - y)^2$$

Where: SS_y = sum of squares for the y matrix

N = total number of samples in the y matrix

i & j = alleles represented in the population

$\sum(y_{ij} - \bar{y})^2$ = the square of how much each value in the y matrix deviates from the mean

$$SP_{xy} = \sum_{i \neq j}^{N} (x_{ij} - \bar{x})(y_{ij} - \bar{y})$$

Where: SP_{xy} = sum of the cross products of the deviations of the matrices x and y from the mean

$$R_{xy} = \frac{SP_{xy}}{\sqrt{[SS_x \times SS_y]}}$$

Where: R_{xy} = Mantel test derived from the fraction of sum of the deviations' cross products and the square root product of the sum of squares for the x and y matrices

The geographic co-ordinates uploaded to GenAlEx were in the form of latitude and longitude in decimal degrees. The distance (D) was then computed in kilometres (km)

$$D = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Where: *xi and yi* are the coordinates for the *i*-th allele and *xj* and *yj* are the coordinates for the *j*-th allele.

5. Spatial autocorrelation analysis using GenAlEx

The formulae for the spatial autocorrelation test are (Smouse & Peakall, 1999):

$$r = \frac{2\sum x_{ij}\, C_{ij}}{\sum x_{ij} C_{ii} + x_{ij} C_{jj}}$$

Where: C_{ij}, C_{ii} and C_{jj} = respective elements of the covariance matrix obtained from the ith and jth alleles
x_{ij} = respective elements of the x matrix obtained from the ith and jth alleles.

The covariance matrix, C_{ij} is computed using the formula:

$$C_{ij} = \frac{1}{2}\left[- {d_{ij}}^2 + \frac{1}{N}\left(\sum_{i=1}^{N} {d_{ij}}^2 + \sum_{j=1}^{N} {d_{ij}}^2 \right) - \frac{1}{N^2} \sum_{i \neq j}^{N} {d_{ij}}^2 \right]$$

Where: The $X_{ij}C_{ij}$ value is the covariance matrix value for pairwise comparisons of two samples – i.e. the element in the matrix with i = 1 and j = 2. $X_{ij}C_{ii}$ is the covariance matrix value of the diagonal element with i = 1 and j=1. $X_{ij}C_{jj}$ is the covariance matrix values of the diagonal element with i=2 and j=2.

APPENDIX H: ALLELE FREQUENCIES

Allele frequency estimates computed per country and per locus including weighted and unweighted by sample sizes values for both SNPs and microsatellites data. The numbers of samples used per country vary due to the existence of missing data in some of the loci. The No. denotes the number of alleles observed in each locus in the case of microsatellites.

Poly - α					
	Cameroon	**Mali**	**Kenya**	**Weighted**	**Unweighted**
No.	**12**	**42**	**23**		
1	0.083	0	0	0.013	0.028
2	0	0	0.087	0.026	0.029
3	0	0.071	0.087	0.065	0.053
4	0.083	0.071	0	0.052	0.052
5	0	0.119	0.109	0.097	0.076
6	0.25	0.143	0.239	0.188	0.211
7	0.167	0.119	0.217	0.156	0.168
8	0	0.119	0	0.065	0.04
9	0	0.071	0.087	0.065	0.053
10	0.083	0.024	0	0.026	0.036
11	0.333	0.071	0	0.091	0.135
12	0	0.048	0.043	0.039	0.03
13	0	0.119	0.043	0.078	0.054
14	0	0	0.087	0.026	0.029
15	0	0.024	0	0.013	0.008

TA81					
	Cameroon	**Mali**	**Kenya**	**Weighted**	**Unweighted**
No.	**12**	**44**	**23**		
1	0.167	0.057	0	0.057	0.074
2	0.25	0.091	0.435	0.215	0.259
3	0	0.364	0.13	0.241	0.165
4	0.25	0.193	0.087	0.171	0.177
5	0.083	0.045	0.174	0.089	0.101
6	0.083	0.205	0.087	0.152	0.125
7	0	0.045	0.087	0.051	0.044
8	0.167	0	0	0.025	0.056

TA17					
	Cameroon	**Mali**	**Kenya**	**Weighted**	**Unweighted**
No.	**12**	**42**	**23**		
1	0	0.024	0	0.013	0.008
2	0.083	0	0	0.013	0.028
3	0	0.095	0.217	0.117	0.104
4	0	0.048	0.087	0.052	0.045
5	0	0.024	0	0.013	0.008
6	0.167	0.19	0.043	0.143	0.134
7	0.167	0	0	0.026	0.056
8	0	0	0.087	0.026	0.029
9	0	0.31	0.348	0.273	0.219
10	0	0.071	0	0.039	0.024
11	0	0.024	0	0.013	0.008
12	0.083	0	0	0.013	0.028
13	0.167	0.19	0.217	0.195	0.192
14	0.167	0	0	0.026	0.056
15	0.083	0.024	0	0.026	0.036
16	0.083	0	0	0.013	0.028

TA53					
	Cameroon	**Mali**	**Kenya**	**Weighted**	**Unweighted**
No.	**12**	**42**	**22**		
1	0	0.048	0	0.026	0.016
2	0.167	0.179	0	0.125	0.115
3	0.667	0.702	0.591	0.664	0.653
4	0	0.048	0.318	0.118	0.122
5	0	0.024	0	0.013	0.008
6	0.167	0	0.091	0.053	0.086

PFPK2					
	Cameroon	**Mali**	**Kenya**	**Weighted**	**Unweighted**
No.	**12**	**39**	**23**		
1	0.25	0.154	0.043	0.135	0.149
2	0	0.205	0.261	0.189	0.155
3	0.167	0.256	0.348	0.27	0.257
4	0.167	0.154	0	0.108	0.107
5	0.167	0.051	0.217	0.122	0.145
6	0	0.051	0.13	0.068	0.061
7	0.083	0.026	0	0.027	0.036
8	0	0.051	0	0.027	0.017
9	0.167	0	0	0.027	0.056
10	0	0.051	0	0.027	0.017

	Cameroon	Mali	Kenya	All_W	All_UW
Pf-01-101502					
	12	**47**	**23**		
A	0.917	0.915	1	0.939	0.944
C	0	0.085	0	0.049	0.028
T	0.083	0	0	0.012	0.028
Pf-01-205304					
	12	**49**	**23**		
A	0	0.041	0	0.024	0.014
C	0.5	0.776	0.478	0.655	0.585
G	0.333	0.102	0.261	0.179	0.232
T	0.167	0.082	0.261	0.143	0.17
Pf-01-411796					
	12	**49**	**23**		
A	1	0.98	1	0.988	0.993
C	0	0.02	0	0.012	0.007
Pf-03138-7					
	12	**49**	**23**		
A	0	0	0.174	0.048	0.058
G	0.833	0.653	0.478	0.631	0.655
T	0.167	0.347	0.348	0.321	0.287
Pf-03-350224					
	12	**49**	**23**		
A	0	0.02	0	0.012	0.007
C	0.083	0.143	0.13	0.131	0.119
G	0.667	0.388	0.435	0.44	0.496
T	0.25	0.449	0.435	0.417	0.378
Pf-03-613146					
	12	**49**	**23**		
C	1	1	0.652	0.905	0.884
G	0	0	0.217	0.06	0.072
T	0	0	0.13	0.036	0.043
Pf-03-842383					
	12	**49**	**23**		
A	1	1	1	1	1
Pf-03-967838					
	12	**49**	**21**		
A	1	1	1	1	1

	Cameroon	Mali	Kenya	All_W	All_UW
Pf-10658467					
	12	**49**	**23**		
A	1	0.918	0.826	0.905	0.915
C	0	0	0.087	0.024	0.029
G	0	0.082	0.087	0.071	0.056
Pf-10-992433					
	12	**47**	**23**		
A	0.75	0.596	0.522	0.598	0.622
C	0	0.021	0	0.012	0.007
G	0.25	0.383	0.478	0.39	0.37
Pf-14-279668					
	12	**49**	**23**		
A	0.917	0.816	0.522	0.75	0.752
G	0.083	0.061	0.304	0.131	0.15
T	0	0.122	0.174	0.119	0.099
Pf-1710196					
	12	**49**	**23**		
A	0.25	0	0	0.036	0.083
G	0.583	0.408	0.435	0.44	0.475
T	0.167	0.592	0.565	0.524	0.441